FORECASTING INSTABILITY INDICATORS IN THE HORN OF AFRICA REGION

THESIS

Bryan R. Tannehill

AFIT/GOR/ENS/08-21

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GOR/ENS/08-21

FORECASTING INSTABILITY INDICATORS IN THE HORN OF AFRICA REGION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Bryan R. Tannehill, MS

March 2008

AFIT/GOR/ENS/08-21

FORECASTING INSTABILITY INDICATORS IN THE HORN OF AFRICA REGION

Bryan R. Tannehill, MS

Approved:

Dr. Richard F. Deckro (Chairman) date

Dr. Kenneth W. Bauer (Member) date

Abstract

The forecasting of state failure and the associated indicators has been a topic of great interest to a number of different agencies. USAid, CENTCOM, the World Bank, the Center for Army Analyses, and others have all examined the subject based on their own specific objectives. Whether the goal is denying terrorists space in which to operate, deciding how to pre-position materials in anticipation of unrest, stabilizing foreign markets and trade, or preventing or mitigating humanitarian disasters, man made or otherwise, this topic has been of interest for over a decade.

The Horn of Africa has been one of the least stable regions in the world over the past three decades, and a continual source of humanitarian crises as well as terrorist activity. Some of the initial modeling of instability was done in response to crises in the Horn of Africa, but research is ongoing. Current models forecasting instability suffer from lack of lead time, subjective predictions, and lack of specificity. The models demonstrated in this study provide 4 year forecasts of battle deaths per capita, refugees per capita, genocide, and undernourishment for Djibouti, Ethiopia, Eritrea, Kenya, Somalia, Sudan, and Yemen. This thesis used principal component analysis, canonical correlation, ordinary least squares regression, logistic regression, and discriminant analysis to develop models of each instability indicator using 54 variables covering 32 years of observations. The key variables within each model are identified, and the accuracy of each model is compared with current models.

Acknowledgments

I would first like to thank various experts in the field who were kind enough to entertain my questions and requests for data regarding their work: Dr. Gurr, Dr. O'Brien, Dr. Marshall, Dr. Woodward, Dr. Wilkenfeld, Dr. Ulfelder, Dr. King, Dr. Pate, and Dr. Hoeffler. Thank you to Dr. Kitashvili and Dr. Walker at USAid for your unwavering support and assistance in tracking down people and information. I sincerely hope this study helps you.

To Dr. Bauer, my reader and multivariate guru, thanks for 685 and 785, even if my questions drove you nuts. To the two Todds, thanks for taking on big group projects with me so I could better understand Nate's dataset.

To Dr. Deckro, my advisor, thank you for your help, patience, and most of all forbearance with my quixotic quest to finish a thesis and find a job.

To my daughters; no matter how long or stressful my day was, coming home and reading bedtime stories to you was always more than made up for it. We'll always have "Harry Potter".

Most of all, I cannot express my gratitude to my wife, who has offered her unconditional support for my work, both here and in the field over the past two and a half years.

Table of Contents

Page

List of Figures

List of Tables

PREDICTING FAILING STATE INDICATORS IN THE HORN OF AFRICA REGION

Introduction

1.1 Background

The objective of this thesis was to provide accurate predictive models of instability indicators in Horn of Africa region. This was done using existing and imputed data to develop substantive models of four continuous instability indicating variables. The application of this is to provide early warning prior to a country experiencing failure conditions. Such early warning would allow various agencies to take immediate preventative measures to prevent a crisis, deny terrorist safe zones, and save money via prophylactic actions. There are strong reasons why such actions are in the self interest to the US government.

Involvement in peacekeeping or stability operations has been unpopular and expensive for the U.S. since at least the collapse of the Soviet Union. The U.S. and United Nations mission to Somalia from December 1992 until March 1995 was unsuccessful, and became unpopular after the U.S. sustained significant casualties. The peacekeeping missions in the Former Republic of Yugoslavia have been ongoing since 1995, despite initial intentions for U.S. forces to be there for a limited time. U.S. operations in Iraq have suffered the same problems, and represent a significant drain on the budget and U.S. military readiness. In addition, the operations in Iraq are increasingly unpopular. Indeed, current research indicates these effects are almost inevitable (Artelli, 2007).

As a result the U.S. government and the DoD are trying to take a more proactive approach to head off conflict before it actually occurs. As part of the Somalia aftermath the Political Instability Task Force (formerly the State Failure Task Force) was formed in 1994 to develop models of instability. (Goldstone, *et al*, 2005, 3) In September 2001 the Center for Army Analysis (CAA) released a global predictive model called Analyzing Complex Threats for Operations and Readiness (ACTOR). USAid has developed methods for assessing stability over the past decade as a way to help focus their operations. All of these efforts were aimed at developing strategies for deterring violent conflict, and reflect a desire to avoid peacekeeping operations. To whit, DoD Directive 3000.05 delineates this philosophy in section 4.3:

> "Many stability operations tasks are best performed by indigenous, foreign or U.S. civilian professionals. Nonetheless, U.S. military forces shall be prepared to perform all tasks necessary to establish or maintain order when civilians cannot do so. Successfully performing such tasks can help secure a lasting peace and facilitate the timely withdrawal of U.S. and foreign forces." (DoDD 3000.05, 2005:2)

The threat of terrorism has exacerbated the desire to better understand stability and its correlates. The U.S. government has stood up AFRICACOM under the DoD, and has put out a comprehensive request for information to assist them in its establishment. There is particular interest in the relationship between terror, instability, and the Horn of Africa region. The Combating Terrorism Center at West Point has studied the attempts by Al Qa'ida to establish operations there since the collapse of the Somali government in 1992. (Harmony Project, 2006:3). Others have suggested the link between failed states and terrorist organizations looking both for recruits and a location from which to base their operations (Forest, 2006: 17-18). The conditions endemic to failed states, such as poverty, unemployment, violence, and lack of legal authority would seem to offer both to terrorist organizations. Based on recent history in Iraq, US leadership would also be less willing to invade another nation, or set up peacekeeping operations in a failed state. (Takeyh and Gvosdev, 2002: 98-101)

Given the current climate towards peacekeeping and the limited funds available, U.S. and allied efforts to stabilize other countries must be focused and efficient. Regardless of the view of the actual relationship between terrorism and stability, both models show a connection. This thesis develops a quantitative model to describe which countries are at greatest risk of experiencing instability events in the Horn of Africa with in four years in order to have time to intervene when the most effect can be gained at the least cost.

1.2 Problem Statement

Numerous government and non-government agencies track data from countries around the world, as well as the Horn of Africa region. Each has their own agenda, and their own issues they wish to push to the forefront. Different groups focus on corruption, freedom, human rights, economics, sociological effects, political rights, and other aspects they have particular interest in. Many even attempt to address stability directly. Due to the endemic poverty and chaotic history of the region, the data can be sparse, or non-existent. However, several groups and researchers have looked at stability globally and within the Horn of Africa. Often these studies suffer from lack of mathematical or statistical rigor and may rely on subject matter experts whose opinions may be difficult to repeat, or are limited by the breadth of their data.

The US allocated $145 Billion for the 2008 fiscal year for anti-terrorism and stability operations in Afghanistan and Iraq. Compare this with the projected $226 Billion dollar deficit and the $36 Billion combined budget request for the Department of State, USAID international affairs, foreign operations budgets (Congressional Budget

Office, 2007: 1), (USAid, 2007, 1). This seems indicative of the disparity of cost in preventing instability, and of efforts in re-building failed states.

Numerous models exist for predicting and fixing failed states. All suffer variously from subjective or qualitative attempts to define stability, lack of time series analysis, a global focus which may mask some variables as noise, not exploring enough variables, or have been ongoing for over a decade without a conclusion. Others have been limited in scope to examine stability from a single perspective or theory of stability. Many studies simply try to describe stability, and not forecast it. Some of the forecasting models do not provide a large enough lead time to allow longer term preventative actions. Given the time it takes to mount an effort to stabilize a country, once a country has begun to noticeably destabilize, it is often too late to prevent complete or partial collapse (Durch, 2002: 18)

This research attempted to forecast potential instability using quantitative and objective measures and to determine which factors are conducive to stability far enough in advance to be useful to planners. Rigorous application of quantitative methods provided a model using existing ongoing data series which should allow focused preventative measures to be taken far enough in advance to be a cost effective tool in avoiding state failure.

1.3 Approach and Methodology

This paper proposes to use several regression techniques and multivariate analysis to predict measures of stability in the Horn of Africa region and to identify the most significant variables within the models. Other research has used different methods such as Discriminant Analysis (DA) and Factor Analysis (FA) (Nysether, 2006: 1-4), used smaller datasets (CAA-R-01-59, 2001: 17), subjective or qualitative response data, or have been global studies which compare data from different regions. Many of them focus on specific academic areas of interest, such as econometric or ethnic causes to the exclusion of a more holistic approach (Collier and Hoeffler, 2001: 2). Some similar research has been done by the Political Instability Task Force. However they have not published final results of their study, nor have they looked specifically at the Horn of Africa region. Their results were also based on a global model (Goldstone, Bates, *et al.*, 2005: 23). In addition, their lead time of two years in their predictive model may be insufficient to mobilize an effective response to some more systemic problems.

Multivariate analysis and regression are used to identify the important variables and variable interactions in a model that predicts stability four years in advance. Stability is measured by the continuous, objective, and quantitative annual data that reflects events that place a nation in a situation which the international community is unwilling to help it out of. These stability indicators are: battle deaths per capita, refugees per capita, genocide and politicide deaths, and undernourishment as a proxy for starvation deaths per capita. The final result are forecasting models for each of the four instability indicators, which allows agencies to identify those countries most at risk of experiencing specific instability events within the next 4 years that will prevent further intervention.

1.4. Research Scope

The goal of this thesis was to generate accurate models of instability indicators and identify the factors and variables that are significant within the models of instability conditions in the Horn of Africa region with a four year lead time. The first hypothesis is that the response data mentioned earlier can be predicted accurately with unclassified time series data. The second hypothesis is that significant variables within the models can be identified, and potentially provide insight into the nature of instability predictors. The significant variables are not necessarily causes, since the research only shows lead and correlation, not causality. Specific actions taken based on results of the model are left to the individual agencies.

The Center for Army Analysis describes instability and violent conflict in terms of an "oily rag" analogy.

> "…then these factors may serve as "oily rags" for a potentially combustible situation. The oilier the rags, the more likely a single spark (i.e., riot, natural disaster, or assassination) could produce an explosive situation. Conversely, the better performing a country is with respect to these factors, the less oily those rags, the more likely it can marshal the will and capacity to withstand a series of sparks or shocks to the system imploding under the weight of the event(s)." (CAA-R-01-59, 2001: 17)

This thesis subscribes to the CAA interpretation of instability, and seeks to put stability in relative terms within the Horn of Africa region. Predicting events such as assassinations or natural disasters is beyond the scope of this research. It is instead trying to define how "oily" the pile of rags may be.

Data in this research is limited to unclassified, open source information only. This is intended to enhance reproducibility, as well as keep the document itself open to examination by the academic community at large. This is not to say that data will not be imputed or forecast in order to create complete rectangular data matrices. Multiple techniques for filling in missing data are used. This is very often the case with survey data, and especially so when dealing with data on some of the poorest and most chaotic countries on earth (Allison, 2001:1)

There are innumerable statistical ways to describe countries. The World Bank maintains a database of over 200 variables, and the United Nations Common Database has 430. Other agencies such as the CIA, US census bureau, World Health Organization, and others maintain or have had databases of similar size. The data series used in this research will not be comprehensive, but limited to those variables available suggested by experts or previous researchers in the field as significant. One goal is to identify which of those variables proposed and examined contribute significantly to the models. The methodologies in this document will allow any variable proposed in the future as an indicator of stability to be tested, and incorporated into the model if significant.

1.5 Assumptions

One underlying assumption of this research is that there is a strong enough relationship between the variables and the response data that a model can be constructed which gives error rates lower than current models while providing longer lead times. This assumes that there is some pattern to war, hunger, genocide, and refugees in the Horn of Africa region. It also assumes that some data is available, and has been collected in a somewhat uniform manner. Where possible, data was collected from single sources, however exceptions are noted. Some data aggregation was used in order to facilitate data imputation and interpolation.

This document also examines both continuous and discrete models of stability indicators. Previous attempts using neural networks, FA, and DA have focused on classification among a small set of defining states as failing or not. (Nysether, 2006: 4) Based on some of the assumptions of the techniques used, transformations of the data to normalize their distributions were implemented as well, making interpretation more difficult.

1.6 Overview

Chapter 2 provides background on the literature describing causes of failing states, as well as variables found to be significant in previous research. Current research on the impact of failed and near failing states will be discussed. Chapter 3 discusses methodologies for dealing with missing data, and the theory behind the techniques used to develop the models of stability variables. Chapter 4 demonstrates continuous and discrete various models for each type of instability indicator using data from the entire region. The variables significant to each model are discussed. This paper also compares and contrasts with other models by previous researchers. Chapter 5 concludes this thesis, and discusses its findings, importance, and suggestions for future research in the field.

2 Literature Review

2.1 Introduction

This chapter reviews the literature and theory regarding failed states used in this thesis. Some general history of the region is given as background and as a framework to better understand the results. Previous efforts have been made to examine stability, either in the region or on a global basis, each with their own strengths and weaknesses. These efforts are reviewed along with relevant terms and definitions. The last portion of the chapter reviews some of the Operations Research (OR) definitions and techniques employed as part of this study.

The results of previous studies, as well as theories put forward by subject matter experts provided potential variables for a forecasting model of instability. These causes, where possible, are used to determine which variables should be included in the construction of a dataset intended to model and forecast state failure. This document discusses some of the available methods of interpolating, extrapolating, or imputing data to create a completely filled rectangular data matrix of variables. Due to the nature of the countries being studied, dealing with missing and incomplete data is one of the largest and most time consuming challenges.

Previous models of instability have used Ordinary Least Squares (OLS), logistic regression, Factor Analysis (FA), Discriminant Analysis (DA), and Neural Networks (NN). Their results are also reviewed. OLS regression is the primary focus of discussed analysis techniques in this chapter.

2.2 Preventing Instability

Governmental stability is of great importance to an array of agencies. The Department of Defense (DoD), USAid, and Department of State (DoS) all have sub-groups dedicated to examining the topic. Former Vice President Al Gore commissioned the Political Instability Task Force in 1995 specifically to address this subject; they continue to conduct research in the area. Civilian agencies such as the Fund for Peace (FfP), United Nations, World Health Organization, and others have developed models to both predict and deal with the consequences of what is regarded as instability.

Obviously, preventing instability is an important goal. More recent studies by the Harmony Project at West Point have indicated that it is perhaps even more important in terms of the war on terror to prevent countries from remaining weak and susceptible to instability. Their case studies of Kenya and Somalia in the 1990's indicate that a weak state is more hospitable to terrorist activity than a failed one (Harmony Project, 2006, 47). The other goal of preventing instability is a more economic one. It is suggested the US is more likely to be involved in "brush wars" attempting to restore stability than it is to be involved in major regional conflicts (Durch, 2002: 8). This study is driven by the belief that an ounce of prevention is worth a pound of cure coupled with the knowledge that once a nation falls into conditions usually identified as state failure it is extremely difficult and expensive to restore it to a semblance of order .

2.2.1 Terms and Definitions

Accelerator or Trigger. Events outside the model which cause a destabilizing feedback loop of violence or unrest. These are factors which are not directly captured in the data, and are "one offs" which lead to greater and greater loss of governmental authority, scope, and range (Gurr & Harff, 1996:47). They can include such things as assassinations, contested elections, natural disasters which expose a regimes inability to provide for its people, enacting new discriminatory policies and laws, and other actions which signal the rapid deterioration of a government. Typically, triggers occur three to six months prior to governmental collapse, and are often not captured using yearly data (Gurr and Harff, 1998: 570).

Armed Conflict. Conflict with weapons between two entities, one of which is the state. This can include invasions, incursions, cross border clashes, and also civil strife with some external dimension (such as the involvement of Sudan in the Chadian/ Libyan conflict in 1988) (Evans, 1993:7).

Civil War or Civil Conflict. Wide scaled armed conflict between the state and internal entities, or in the case of failed states with no central authority, between two communal groups (Weiss & Collins, 1996:217).

CNN-Factor. The CNN-Factor refers to the emotional reaction of the public to media coverage of events or conditions. Debate continues as to how a public's reaction to what they see on TV can influence their government's response to a crisis in another country (Schmid, 1998:11).

Failed State: A failed state is one in which the government cannot provide "positive political goods" to its population. The government cannot in effect provide a benefit to those they rule or represent. They also cannot effectively control their sovereign clamed territory, and there is no clear power in control over the territory claimed. Pakistan and its Waziristan region are a good example of the latter (Rotberg, 2003: 1-2). Somalia exemplifies all of these characteristics.

Fractionalization. The probability that two randomly selected individuals from a population come from the same group. This can be applied to language, ethnicity, religion, or a combination of each. A low score indicates high fractionalization. This Ethno-Linguistic Fractionalization (ELF) score has been found to have a high correlation with per capita GDP (Alesina, *et al*, 2003, 158-159).

Genocide.

> "..involves the promotion, execution, and/or implied consent of sustained policies by governing elites or their agents or in the case of civil war, either of the contending authorities that result in the deaths of a substantial portion of a communal group… In genocides the victimized groups are defined primarily in terms of their communal (ethnolinguistic, religious) characteristics" (Gurr & Harff, 1998: 1).

In the Horn of Africa, the Ethiopian famine of 1984-1985 and the suppression of Black Muslims in Darfur in 2003 through the present may rise to the level of genocide.

Indicators or Predictors. Precursors of instability. These generally fall into three categories: underlying long term conditions, specific situational causes unique to the area, and accelerators or triggers. This document is concerned primarily with long term conditions.

Political Instability. The inherent tendency of a government towards failure. Previous studies have included casualties from civil wars, ethnic wars, genocide, politicide, and adverse regime changes as indicators of this, and failure (Goldstone, *et al*, 2005: 4). This study uses refugees in place of adverse regime change to obtain a more quantitative measure of the worsening of government policies and decreasing ability to provide political goods to the populace.

Politicide. Similar to genocide, except the targeted group is typically defined by its opposition to the regime in power (Gurr & Harff, 2001: 1).

Refugee. A person who has left their country of citizenship due to fear of persecution or great and grievous bodily harm due to ethnicity, race, religion, or political affiliation and is unwilling to return based on their inability to entrust their safety and rights to government of their country of origin. Recent efforts at the UN have attempted to expand this to include discrimination based on gender. (UNCHR, 1951: 16)

Stability Operations. All military and civilian activities designed to strengthen weak or failing states, as well as restore order in failed states (DoDD 3000.05, 2005:2). One of Durch's central arguments is that stability operations are critical, but getting in sooner rather than later will significantly reduce long term costs. (Durch, 2002: 18)

Severe Malnutrition or Severe Wasting. A z-score of more than -3 standard deviations from the norm on the Body Mass Index (BMI) scale (UNFAO, 1995: 1)

Weak or Failing State. A state in which its populace derives only marginal gain or a large portion of the populace derives none at all from its government. The government is still functional, but with severe limitations. This is different from the anarchy of a failed state, such as Somalia, where there is no identifiable governmental power, or that power has no impact outside a very small area. According to recent theory based on analysis of al-Qaeda operations in the Horn of Africa in the 1990's, weak or failing states offer a better breeding ground for terrorist organizations and operations than failed states. (Harmony Project, 2006, iii)

2.2.2 Current Models and Theories

The subject of stability in the Horn of Africa has been an important topic to the world community for decades. The famine deaths in Ethiopia in 1984 and 1985 shocked the world as much because of suffering, as because the overwhelming supply of aid materials were left to rot on the docks. Suggested explanations for this failure include poor infrastructure, an ineffectual government or, worst of all, an Ethiopian government deliberately trying to reduce the population of an area that was in opposition to the government (BBC News, 2000). The efforts of the US government in Somalia as part of Operation Restore Hope and the UN mission to Somalia from 1992 through 1995, as well as their subsequent failure, illustrated the consequences of allowing a state to fail that completely. It also highlighted the immense difficulty of restoring a nation with poor infrastructure, fractionalization, health, nutrition, no functioning security apparatus, and no functioning judiciary (Allard, 1995: 87). More recently, the humanitarian disaster in Darfur has attracted significant attention, although even the repeated stories in the press were insufficient to motivate governments to intervene due to political and economic issues. Other than calls for calm, no direct action by foreign or international agencies to stem the recent tide of violence in Kenya has taken place.

A number of authors, government organizations, and independent groups have tackled the subject of instability, with the Horn of Africa region frequently used for case studies to provide the data modeling failure conditions. This section discusses a few of these studies, and how this paper deviates from or improves upon them.

One of the most influential models of instability in current use is an econometric model first presented in 1998 by Collier and Hoeffler as a project for the World Bank (Collier and Hoeffler, 1998: 1). Their model from "On Economic Causes of Civil War" first suggested that people, as a group within a country, make rational decisions based on an intrinsic risk reward calculation. They developed a formula that shows the decision to embark on a civil war by rebels can be expressed as:

$$W = 1 \text{ if } U_w > 0 \text{ else } W = 0$$

where W =1 is war and W = 0 is peace and U is the rebel utility function (Collier and Hoeffler, 1998: 2). This function represents whether or not the expected value of a revolution is of benefit to a group as whole. War can cause benefit to a group, such as greater economic equality, greater political representation, or freedom from an oppressive regime. It has costs in terms of economics and lives as well, along with the risk of losing, and being wiped out as a group. Given this, a utility function was defined as follows:

$$U_w = \int_{t=D}^{\infty} \frac{p(T) \bullet G(T,P)}{(1+r)^t} dt - \int_{t=0}^{t=D} \frac{(f(Y)+C)}{(1+r)^t} dt$$

where p = the probability of victory

T = the taxable capacity of the economy

G = gain conditional upon victory

P = the size of the population

D = the expected duration of warfare

Y = per capita income

C = coordination costs

R = the discount rate

This model makes the assumption of perfect knowledge, as well as making an assumption that groups are rational actors that are primarily motivated by personal economic gain. Factors such as grievances, language, ethnicity, and religion as causes are not found to be contributing variables. (Collier and Hoeffler, 1998: 3-4)

Later research by Collier and Hoeffler using a logit regression economic and social data further suggested that political and social factors were less significant than economic ones. They conclude that the factors which contribute the most to their regression model are primary commodity exports as a percent of GDP, GDP per capita, rate of GDP change, and secondary education (as a proxy for job opportunity cost). Their model also only uses total combat deaths as a response (Collier and Hoeffler, 2001, 8-9).

Primary commodity exports as a percentage of Gross Domestic Product (GDP) has a non-monotonic effect on their model, meaning that the variable does not have a uniformly increasing or decreasing effect on the overall model.

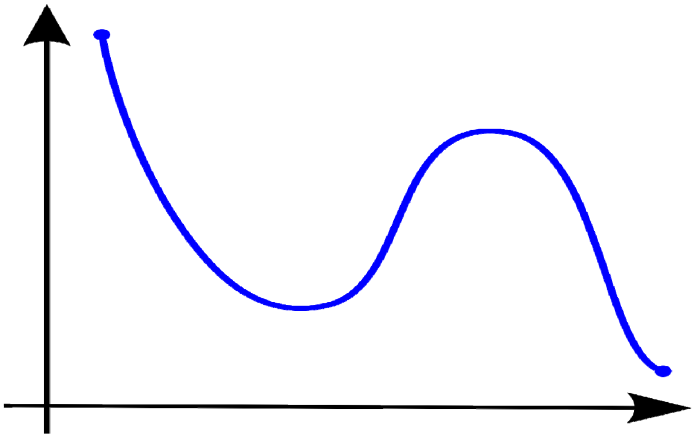


Fig 2-1. Non-Monotonic Function

Countries that are not reliant on any one product tend to be more stable than a country that is heavily reliant on a single one. The peak of instability effect is reached at a level of approximately 26% of GDP. This seems to be true regardless of what the product is, with only oil breaking out as being different, although not substantially so. They examined their findings on a global scale, and also compared their global model's applicability to Africa and Sub-Saharan Africa with good results (Collier and Hoeffler, 2002: 16-17).

Collier and Hoeffler's work has been cited heavily by the World Bank, the UN, and USAid for their own efforts in stabilization operations. However, their model is not the only one that has been developed. Another model that has had significant impact on this research was developed by the Center for Army Analysis (CAA) in 1999. The ACTOR model was intended to allow the Army to better predict where they would deploy to within the next 15 years in order to better pre-position themselves. Their variables were based on assessment of internal state structure, which is similar to Collier and Hoeffler's approach, and a tactic this study has used.

The CAA database was extensive, extending from 1945 through 1999 on 12 variables. Their response variable for conflict was the somewhat subjective KOSIMO project database (Pfetsch and Rohloff, 2007: 380), which rated conflict on a zero to four scale. Based on previous research by the CAA, the variables they identified as potentially significant were percent of national history spent in conflict (years at war divided by years available), infant mortality rate, trade openness, youth bulge, the Freedom House civil liberties index, life expectancy, the Freedom House political rights index, the Polity 98 democracy index, religious diversity, caloric intake per capita per day, GDP per capita, and ethnic diversity. This study took a more holistic approach than Collier and Hoeffler and addressed social, political, and economic factors, and deemed all of them significant. The definitions of variables used in this study are provided in Appendix C.

The CAA tested a number of different mathematical techniques to try to sort countries into the 5 levels in the KOSIMO classification system. Logistic regression, classification and regression trees, temporal decision trees, and neural networks were all tested for their predictive abilities on historical data. In the end, they developed an algorithm called Fuzzy Analysis of Statistical Evidence (FASE) that is a hybrid of several other techniques from statistics, fuzzy logic, and possibility theory. (CAA, 2001: 14)

One difficulty presented by their study is that much of the data showed non-monotonic qualities when comparing it with likelihoods of instability.

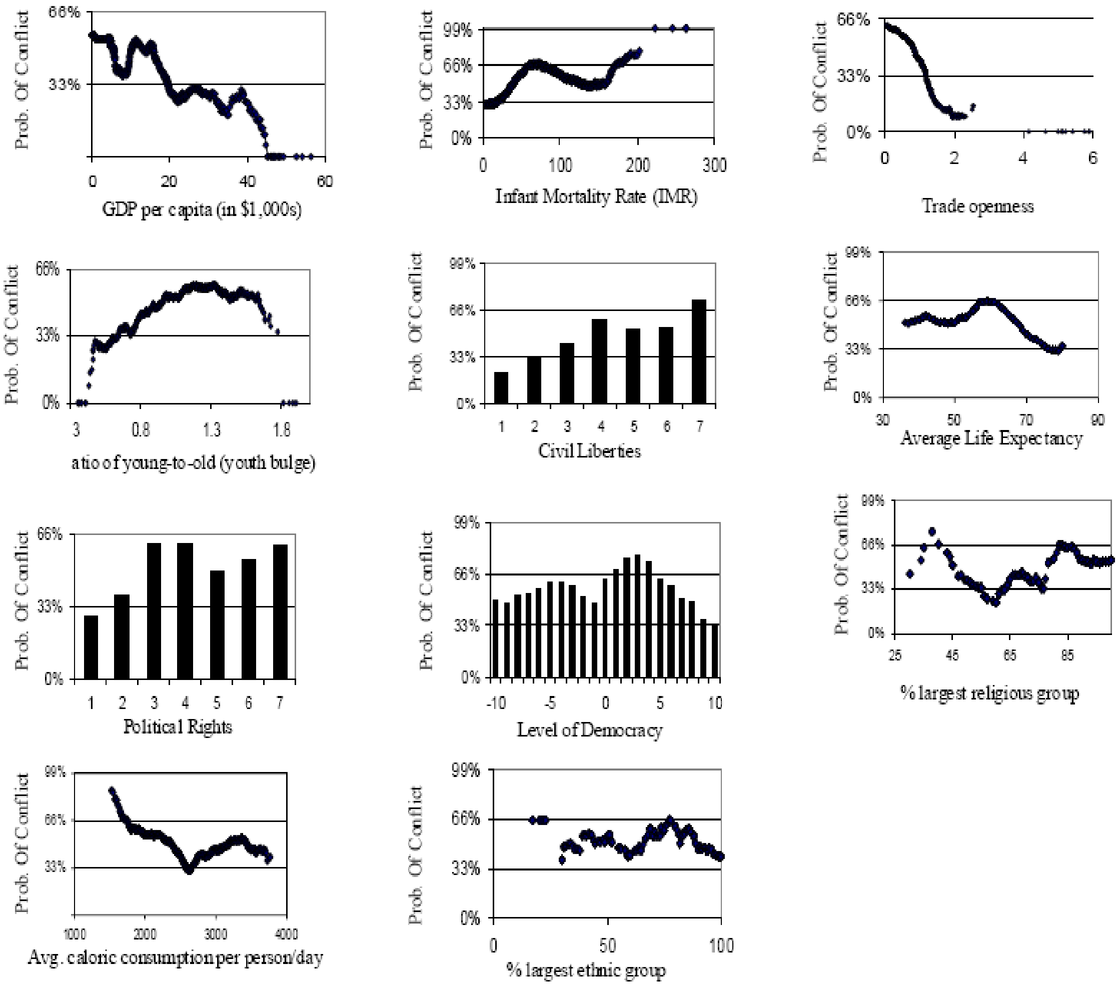


Fig. 2-2. CAA Variable Effects on Probability of Conflict
(O'Brien, 2002: 14)

Of note among the strongly non-monotonic relationships is that the most fragile societies tend to be neither democratic nor authoritarian with Polity score of 0. Largest ethnic group had the least effect on the model, while trade openness looked to be an almost monotonic S-curve. Of note is that GDP per capita, while not monotonic, appears strongly correlated again in the FASE model, as shown in the Collier and Hoeffler model.

One of the first attempts at modeling stability came from the State Failure Task Force (now called Political Instability Task Force (PITF)). They were commissioned in 1994 in the aftermath of Somalia by then Vice-President Gore and the CIA to develop a forecasting model to predict which countries were most likely to fail on a relative scale (Esty, *et al*, 1995, 3-5). They attempted to model adverse regime changes, ethnic wars, revolutionary wars, and genocides / politicides. (Goldstone, *et al*, 2005: 5) Over a thousand variables made up of time series data was tested, and a number of different techniques for classifying states were used including logistic regression, neural networks, Markov processes, and history models. The original intent of the project was to use regression to simply reduce the number of variables in order to simplify the problem being fed into a neural network. However, logistic regression ended up consistently out performing the other techniques. Prediction was done by regressing data from two years prior to instability events. Their findings suggested that their model was universal, applicable to any sub-region and even across cultures and religions (Gurr, *et al*, 2005: 11-12).

Since 1995 the group has released 5 updates on their project. Their initial intentions of using regression to identify significant variables for use in neural networks eventually became to find variables to test using step wise logistic regression in using the case-control method. Eventually, they concluded that it was more rewarding to look at what makes a state stable, rather than unstable. To use the oily-rag analogy, their results did not measure how much oil was in the rags, but how much fire retardant. Their eventual model had only four variables (regime type, infant mortality, "bad neighborhood", and state sponsored discrimination) (Goldstone, *et al*, 2005: 22). Despite numerous attempts to find a more complicated model that significantly outperformed their simple model, they were unsuccessful.

Some of their results dovetailed with CAA ACTOR model results. The PITF also found the least stable countries were those that could be described as illiberal democracies which allow some freedom, but are not strong enough to fully enforce limits. The most stable regimes were full autocracies where the state exerted overwhelming control of political activities. Surprisingly, none of the econometric factors cited by Collier and Hoeffler and tested by PITF contributed to their model (Goldstone, *et al*, 2005: 20-21). In the end, the contributing factors to their model were infant mortality rate, a "bad neighbor" score indicating if four or more neighboring countries were involve in conflict, regime type based on scores from the Polity IV database, the presence or absence of state led discrimination, and the interaction between regime type and discrimination. This interaction was the most powerful factor model, where illiberal democracies with a high level of political discrimination were 36-60 times more likely to experience instability than full autocracies (Goldstone, et al, 2005: 20).

Of note is the 'bad neighbors" or "bad neighborhoods" score. This idea was also put forward by Dr. William Durch as a potential cause of instability in his briefing to a study group for the Chairman of the Joint Chiefs of Staff (CJCS) in 2002. The idea is simply that if the countries along country A's border are at war, but not with country A, those wars have a destabilizing effect on country A. In the case of PITF, bad neighborhoods were measured as a binary score of 1 when 4 or more neighbors were at war in a given year and 0 when 3 or less were at war.   Four was chosen by PITF as a break point, since it typically indicates that a majority of ones neighbors are involved in armed conflict. Dr. Durch also added water scarcity and stress as a potential cause of instability, along with population density. The latter runs contrary to some social theory regarding stability in the developing world and particularly the Horn of Africa (Jacquin-Berdal, 2002, 12). The CJCS briefing fuses much of the theory put forth in PITF and the Collier Hoeffler (CH) model, arguing that econometric, political, and ethnic factors are significant factors in predicting and understanding stability in the developing world. He shows both the relationship between economic disparity and instability correlation between Polity IV scale data and instability, as well as the correlation between Polity IV scale data and instability (Durch, 2002: 9-10). These can be seen in Figures 2-3 and 2-4.

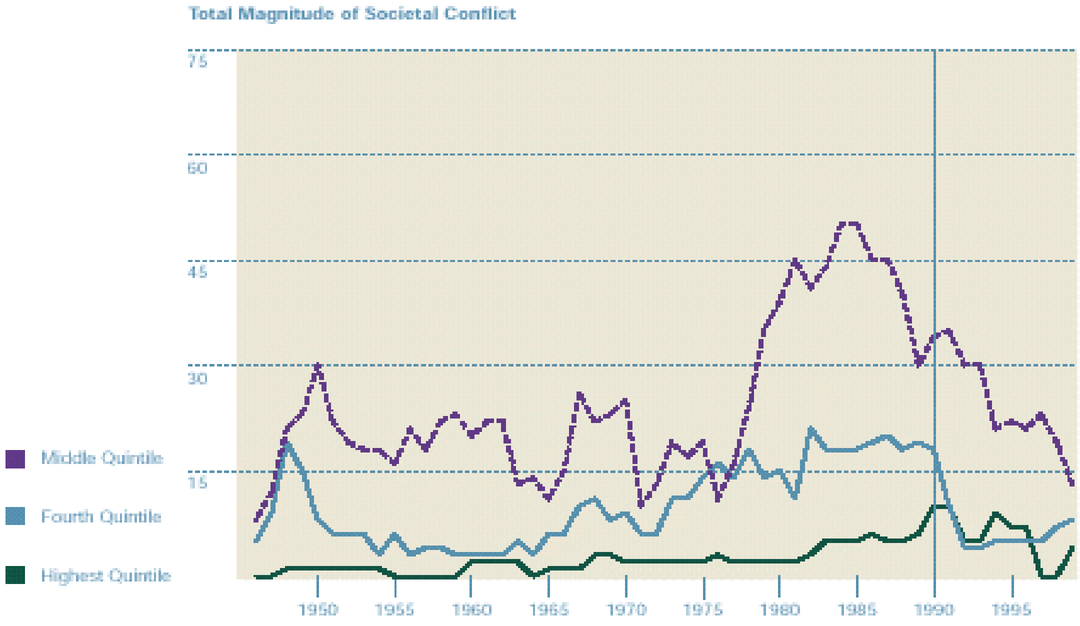


Fig 2-3. Global Warfare By Societal Capacity
(Gurr, Marshall, and Khosla, 2001: 12)

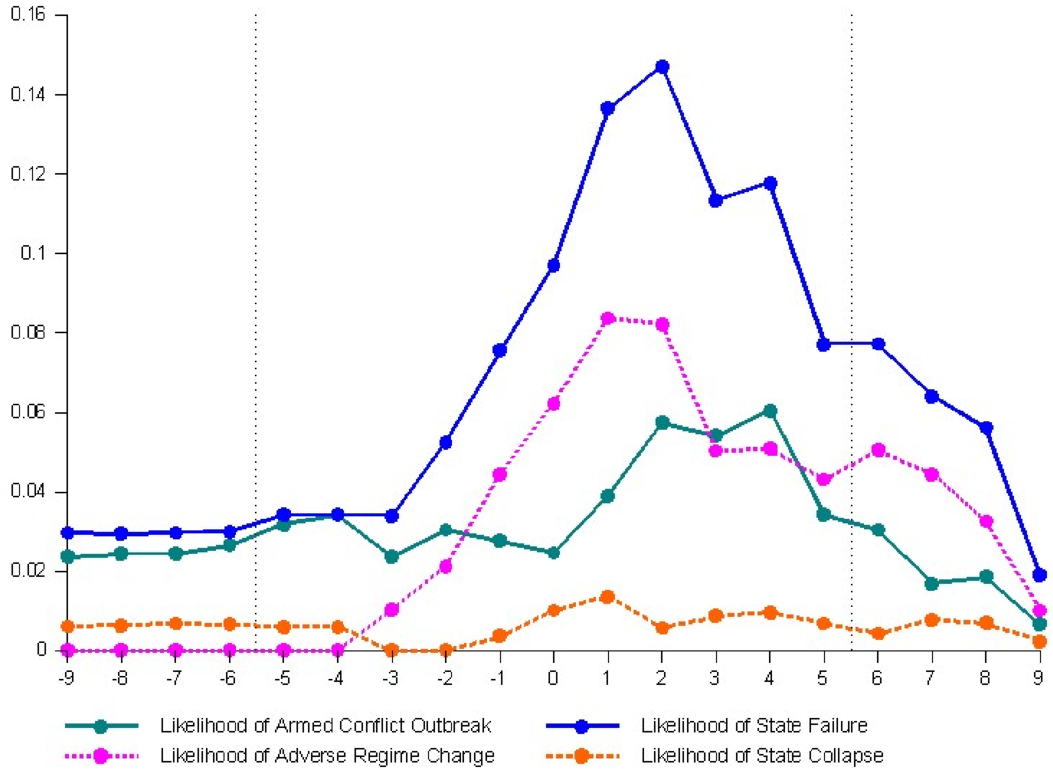


Fig. 2-4. Likelihood of State Failure Events
(Marshall, 2001: 10)

Another significant point Durch makes is giving a theoretical graphical representation of national level destabilization. It agrees with the CAA "oily-rag" description of stability by positing weak or failing countries can remain functional until some event causes them to rapidly fail. Although the graphical depiction (Figure 2-5) does not have an associated time scale, separate research by Gurr and Harff has suggested the triggering event occurs approximately three to six months prior to full destabilization or in this case a state of open warfare (Gurr and Harff, 1998: 570).

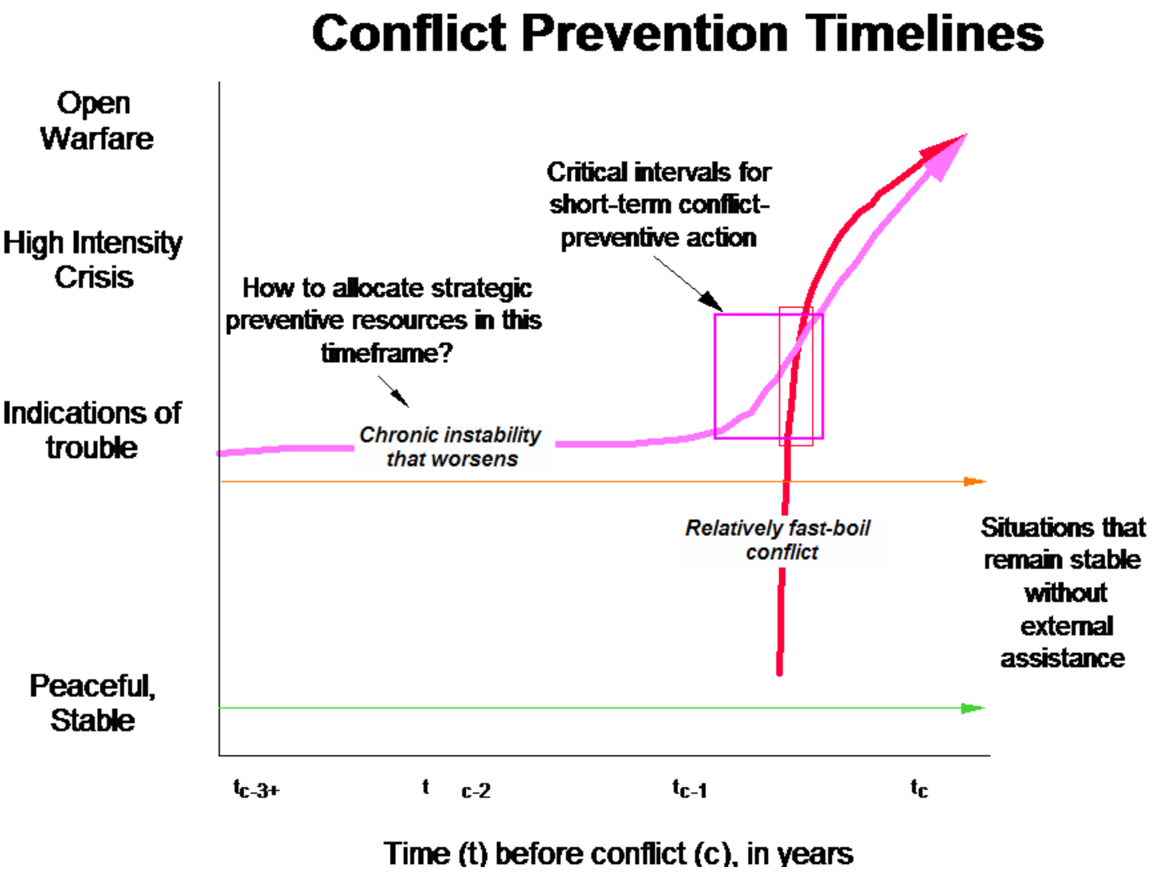


Fig. 2-5 Conflict and Stability Timeline
(Durch, 2002: 18)

One area where there is little consensus is on the issue of ethnicity, language, race, and religion as a factor in stability. Some sociologists have argued that race in the Horn of Africa (HoA) region is primarily a modern Western construct. An example given is Somalia, where there is in fact a fairly homogenous population in terms of ethnicity, but has failed utterly at self government and virtually any measure of stability. One explanation offered is that a sense of nationalism, or a person's primary identification as a member of a nation vice a religious, ethnic, racial, or other group is crucial to stability. Nationalism is at its core derived from the ability to communicate with other people from around the area defined as a nation. Measurable things which theoretically contribute to a sense of nationalism, and hence stability, include a common language, urbanization, literacy, education, communications networks, and a wide spread press. One postulate is that state versus state warfare may actually stabilize a nation by generating a sense of nationalism or national unity (Jacquin-Berdal, 2002: 8-34). Others, such as Gurr, Harff, and the PITF have implicitly indicated that the existence of persecuted minorities who are discriminated against economically and politically is one of the most powerful destabilizing forces in a model of state failure and civil war (Gurr and Harff, 1998: 561).

Several other studies have used other quantitative methods to identify key variables and create stability models based on them. Most recently Capt Nathan Nysether looked at global stability using a database of 167 variables and over 200 countries to build a descriptive, but not forecasting, model of stability. The response variable was Barnett's core, rim, and gap classifications of national economies. The reason for choosing these classifications is Barnett argues that the gap countries which are not fully integrated into the world economy are the ones most likely to fail, and as a result are the ones that will most probably experience humanitarian disasters requiring lengthy and expensive international intervention (Barnett, 2003: 148). Nysether used Factor Analysis (FA) as a method for reducing the number of variables and grouping them, and Discriminant Analysis (DA) to attempt to sort them into the three echelon core, rim, and gap classifications. He also used the subjective four tiered Fund for Peace stability scoring system to test against his core, rim, and gap results (Nysether, 2007, 4-28, 29).

FA when conducted as a varimax rotation tends to group variables together in terms of total variance explained, and a common thread between each member of the group is not often readily apparent. Nysther settled on 10 factors with 32 variables shown in Table 2-1. The variables in bold had a negative effect on stability. Of note is that Nysether's analysis does not constitute a claim of causality, merely one of correlation, nor do any of the other models except the theoretical ones proposed by sociologists and political scientists.

Factor 1: The Big Picture - This factor encompasses the vast majority of variables experts use in determining the overall status of a country, and determining national stability.	
Log(A231+0.001)	Carbon dioxide emissions (CO2), metric tons of CO2 per capita (CDIAC)
Log(A119)	GDP Per Capita
A185	Urban population (% of total)
Log(-A243)	Balance of Payments: imports of goods, free on board, US$ (IMF)
Log(A264)	Number of Recorded Drug Crimes Per 1000 Pop
A192	Children 1 year old immunized against measles, percentage
A175	Ratio of female to male enrollments in tertiary education
A259	Enrolment in total secondary. Public and private. All programs. Total %
A177	Ratio of girls to boys in primary and secondary education (%)
Log(A246+0.001)	Exchange rate, US$ per national currency (IMF)
Log(A167+0.001)	Population growth (annual %)
A120	Political Terror Rating
A257	School age population. Tertiary. Total %
A253	School age population. Primary. Total %
Log(A215)	Tuberculosis death rate per 100,000 population
Log(A193)	Population undernourished, percentage
A190	Children under five mortality rate per 1,000 live births
Log(A174)	Pupil-teacher ratio, primary
Factor 2: Sustainability - This factor seems to capture a country's population and their ability to provide for it. Also included is the Count of Entries, which measures various organizations' ability/desire to collect data on each nation.	
Log(A100)	Population
Log(A155)	Land area (sq. km)
A252	Count of entries in database
Log(-A243)	Balance of Payments: imports of goods, free on board, US$ (IMF)
Log(A248)	Imports of goods and services, current prices
Log(A126)	Aid per capita (current US$)
Factor 3: Women's Rights	
Log(A221+0.001)	Seats held by women in national parliament
Log(A172+0.001)	Proportion of seats held by women in national parliament (%)
Factor 4: Population Growth	
A250	Migration, international net rate per year
Log(A167+0.001)	Population growth (annual %)
Factor 5: Crowdedness	
Log(A166)	Population density (people per sq. km)
Factor 6: Economic Growth	
Log(A144)	GDP per capita growth (annual %)
Factor 7: Crime Rate	
Log(A263)	Number of Recorded Murders Attempted Per 1000 Pop
Factor 8: Openness	
A135	Exports of goods and services (% of GDP)
Log(A153)	International tourism, expenditures (% of total imports)
Factor 9: Displaced Persons	
Log(A118)	Refugees
Factor 10: Military Focus	
Log(A159+0.001)	Military expenditure (% of GDP)

Table 2-1: Factor Analysis Variable Groupings In Nysether Model
(Nysether, 2007: 4-8)

Prior to Nysether's model, another related model of stability operations was made using simulation by Capt Matthew Robbins, USAF. His model separated stability and reconstruction efforts into four fields: security, establishing law and order, maintenance of critical infrastructure, and an effective indigenous government (Robbins, 2005: 6). Of these, security functioned as an enabler of the other three. In contrast to the other models discussed so far, the focus on security led to the inclusion of crime related variables such as access to arms, crime rates, and arrest rates. All of these have been shown to reduce GDP and growth, and as a result make a country more prone to instability (Niskanen, 1994: 1). Robbins also used measures mentioned previously in the literature as potential indicators of stability, such as Collier and Hoeffler's primary commodity exports as percentage of GDP, unemployment of military aged males, fractionalization, discrimination, and military spending as a percent of GDP.

The other aspect of Robbins thesis that has not been addressed is the importance of infrastructure to reconstruction and stability. Previous studies suggest the foremost factor in measuring infrastructure robustness, and the first thing that must be built to develop infrastructure, is roads (Cho, 2007: 15). There are many ways to measure infrastructure in the Horn of Africa, however the data was usually unavailable. Indeed, the Harmony project report on al-Qaeda's (mis)- Adventures in the Horn of Africa shows that the total breakdown of infrastructure in Somalia hampered al-Qaeda's attempts to operate there every bit as much as it did the United Nations' (Combating Terrorism Center, 2006: 22). For this study, some road data was available, although because of missing data it was heavily imputed.

2.3. Missing Data Techniques

This section briefly discusses the techniques used to deal with missing data in the database. It includes definitions and a brief discussion of different techniques, and where they are applicable to the data in this research.

2.3.1. Missing Data Introduction

The problem of missing data is a common one in data intensive real world studies, and very much so when using historical data vice a designed experiment. This is especially true for historical data on countries in the Horn of Africa region, where poverty, instability, and weak governments have traditionally hampered data collection efforts. Poverty also has the effect of rendering these countries less significant economically, lowering the incentive to study and collect data on economies which have little global impact. In this study, time series data for seven countries (Djibouti, Eritrea, Ethiopia, Kenya, Sudan, Somalia, and Yemen) was collected covering the years 1975 through 2006. Much of the early data and some of the most recent data are unavailable or have not been collected. In some cases, no data at all was collected for a particular country for a particular variable.

Generation of a complete rectangular dataset can be vital to further analysis. The mathematical techniques mentioned so far (regression, principle component analysis, FA, and DA), expect complete datasets. There are many techniques available to generate the missing data; some are discussed in this section. However, not all of the approaches have been used on the actual data in this study. See Chapter 3 for implementation details regarding the generation of missing data for this project's dataset.

2.3.2. Terms and Definitions

Autocorrelation. The correlation between two values of a variable (or error terms) spaced *k* time units *t* apart. In simpler terms, the data from 1,2,3, etc… time units ago has a statistically significant effect on the current data. Much of the data in this research is highly autocorrelated, and this is a common relationship observed in economics and engineering. The Durbin-Watson test is commonly used to test for autocorrelation (Montgomery, Peck, and Vining, 2006: 475-476). Autocorrelation between error values in a model significantly alters Ordinary Least Squares (OLS) assumptions.

Correlation. A coefficient between -1 and 1 defined by:

$$\rho = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2}$$

A $\rho > 0$ indicates that as Y_1 increases Y_2 increases. A condition coefficient of -1 or 1 indicates perfect correlation with all points falling on a line with a positive or negative slope depending on the sign of the coefficient. When ρ is zero it implies zero covariance and no correlation (Wackerly, Mendenhall, and Schaeffer, 2002: 250). Generally some correlation is present in real world data. It then becomes a question of the degree of correlation.

Extrapolation, Forecasting, Prediction. These terms describe a statistically driven estimate of data in a time series which extends beyond those already measured which attempts to minimize error by way of an algorithm (Wiener, 1949: 9). Extrapolation and forecasting methods generate values for a set more subject to uncertainty than imputation. The green blocks in Figure 2-6 show where data might be extrapolated.

Imputation. A generic term for filling in missing data with plausible values (Schaefer, 1997: 1). In a multivariate dataset the missing data may be replaced using a regression model from existing data, or from the existing data of the same type. Different methods of imputation are discussed later in section 2.3.5.

Interpolation. Interpolation is the process of using a generated function to fit given data in order to predict data between observations. A function is generated based on the existing points {x}, and the outputs. In the case of the data set for this research the input is the year, and the output is the missing variable data point of interest. For example, if only the GDP per capita of Ethiopia in 1980 was missing for Ethiopia's time series data on GDP per capita, based on all the other points for Ethiopia for years before and after 1980 a function would be generated that will give us an estimate of GDP per capita in 1980 using 1980 as the input variable, and also return the actual data for non-missing points for the years used to build the function (Burden and Faires, 2002: 105). The red blocks in Figure 2-6 show the data points subject to interpolation and the numbers in the white blocks show the real data.

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Literacy Rate	🟩	40.1	🟥	🟥	🟥	🟥	46.4	🟥	49	🟩	🟩	🟩	🟩

Figure 2-6: Potential Interpolation and Extrapolation Blocks

Missing at Random (MAR). Data is MAR in a univariate example if the observed units are a random sub-sample of the sampled units and probability of a data point being missing is independent of its value. If MAR does not hold for a sample, the Missing Data Mechanism (MDM) is non-ignorable and the remaining reduced set is subject to bias (Little and Rubin, 1987: 10) The definitions of MDM and non-ignorable to follow.

Missing Completely at Random (MCAR). Data is MCAR if the probability of a data point being missing is independent of its value and the value of any other variable in the dataset. Thus, MACR implies that no pattern of "missingness" can be observed within the dataset (Little and Rubin, 19887: 14).

Missing Data Mechanism. The set of parameters describing the probability structure of missing data (Allison, 2001: 5)

Ignorable Missing Data Mechanism. Missing data is said to be ignorable if the mechanism that selects data, and missing data, is under the control and understood by the sampler (Little and Rubin, 1987: 9). This is also true in cases where the missing data is MCAR or MAR and the MDM is unrelated to the parameters of interest (Allison, 2001: 5).

Non-Ignorable Missing Data Mechanism. If there is some structure to which pieces of data are missing, then some effort to capture and model this to capture the underlying structure of the data (Allison, 2001: 5).

Stationary Series. A time series is said to be stationary if:

1. There does not appear to be a change in the mean of the series over time.
2. The plotted time series data does not show a change in variance over time.

(Makridakis, Wheelwright, and McGee, 1983, 359).

2.3.3 Interpolation Methods

Interpolation techniques are best suited for data series which one would expect to have very little noise, have strongly auto-correlated values, and in the form of a smooth curve. In fact, some types of splines produce the smoothest possible curve that will fit a set of points. Because of these properties, time series data for large population over a large period of time (one year) tends to produce fairly smooth curves even when the entire data set is available. See Figure 2-8 for an example. Thus, where possible and applicable, splines are used in this study preferentially over extrapolation and imputation, since the latter generally does not take autocorrelation into account when producing fill in data, and the former uses data that is not suitable for interpolation techniques. This also prepares the data for use with the multiple imputation software used to generate the remaining data points (King, 2001: 4).

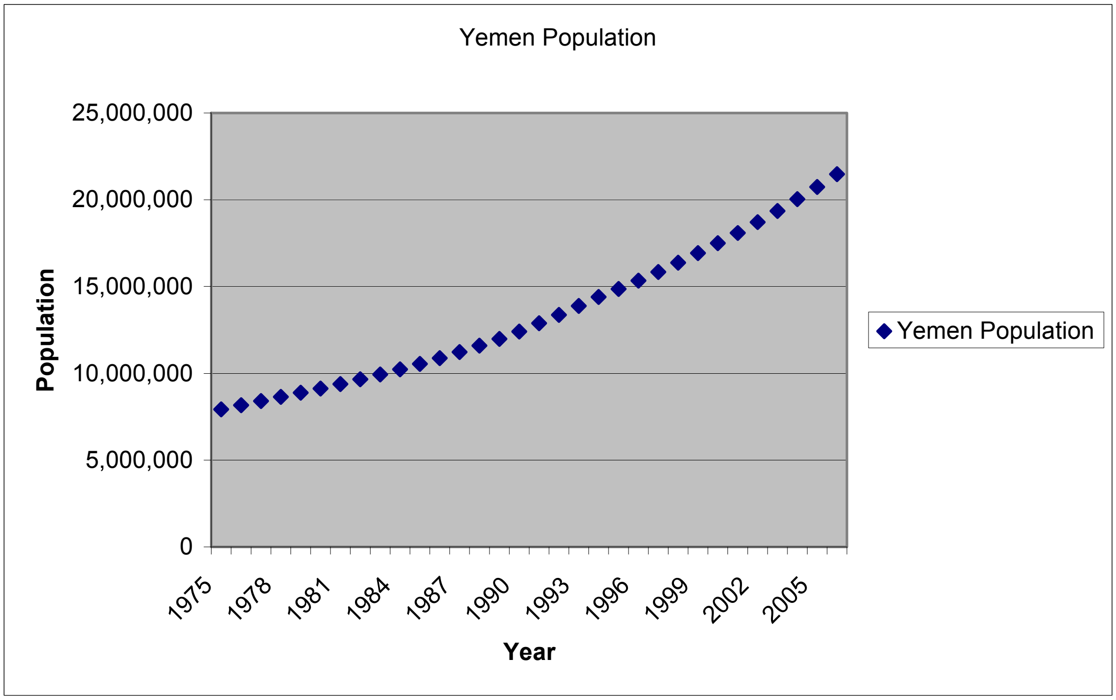


Figure 2-7: Example of Data Suitable for Interpolation

Nearest Neighbor Interpolation. Nearest neighbor interpolation is one of the simplest methods available. The algorithm simply looks for the closest real (measured) data point and imputes it to the value being sought. For example, if one were interested in the infant mortality rate in the Sudan in 1978, but only had data points for 1977 and 1980, the 1978 data point would take on the value from 1977. This creates a piece wise discontinuous data set. The method has the benefit of being simple and requiring minimal computation time, but is not nearly as accurate as other methods available.

Piece-Wise Linear Interpolation. This is a special case of spline interpolation using first order polynomial approximations. In this approach every set of points is splined together using a first order polynomial equation. Suppose, given some x_0, $f(x_0)$, x_1, and $f(x_1)$, the analyst wished to find some value $f(x_2)$ for an x_2 that lies between x_0 and x_1. The equation

$$f(x_2) = f(x_0) + \frac{(x_2 - x_0)(f(x_1) - f(x_0))}{x_1 - x_0}$$

yields any value of x_2 so long as it is between any two other existing values. This approach can be used repeatedly between all existing values to yield a continuous but non-differentiable line, as seen in Figure 2-9 (Burden and Faires, 2001: 109).

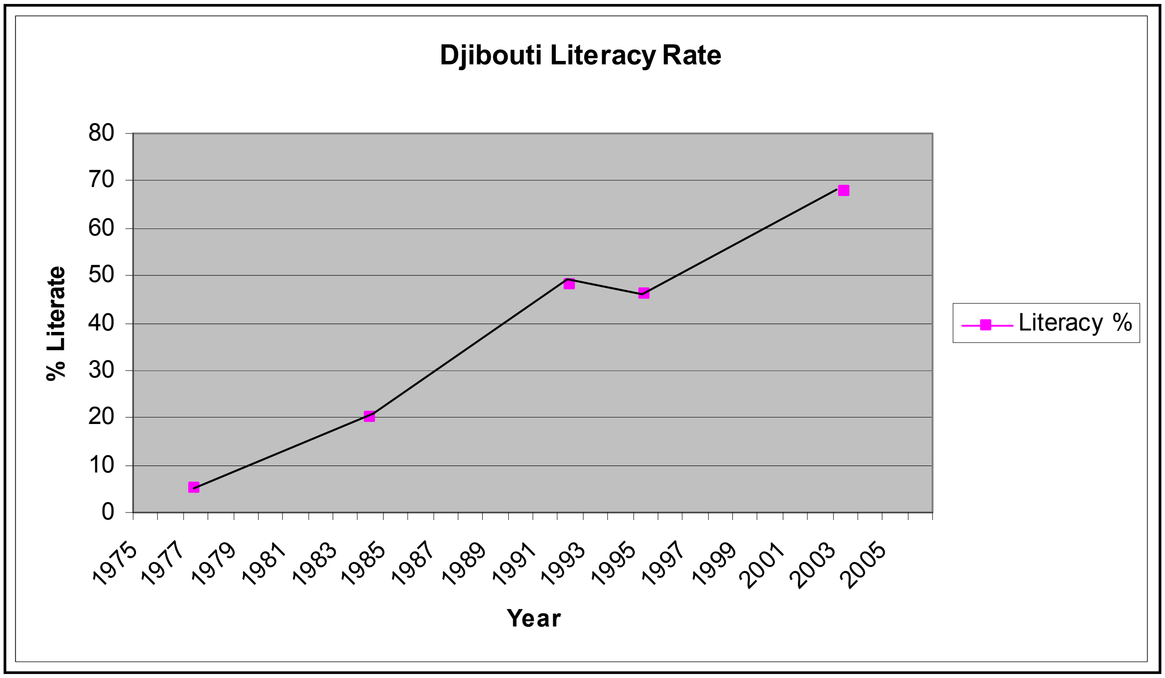


Figure 2-8: Example of Piece Wise Linear Interpolation

(Lagrange) Polynomial Interpolation. If there is *n+ 1 point* to be connected, a polynomial interpolation generates a unique nth degree polynomial *P(x)* which joins the real points and provides an estimate of all points in between the *n+1* knots where

$$P(x) = \sum_{k=0}^{n} f(x_k) L_{n,k}(x)$$

and where $L_{n,\, k}(x)$ for each k = 0, 1,…., *n* is

$$L_{n,k}(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)}$$

(Burden and Faires, 2001:109-110)

Polynomial interpolation algorithms are relatively simple. However, they have some problems with error. Notably as the degree of the interpolating polynomial grows with evenly spaced points, it becomes more and more susceptible to large oscillations or spikes which induce error in the interpolated values. This is known as Runge's phenomenon (Schatzman, 2002: 61). Figure 2-9 shows the results of polynomial interpolation of the data from Figure 2-8. Variable functions, which are not smooth and have many data points, are poor candidates for polynomial interpolation.

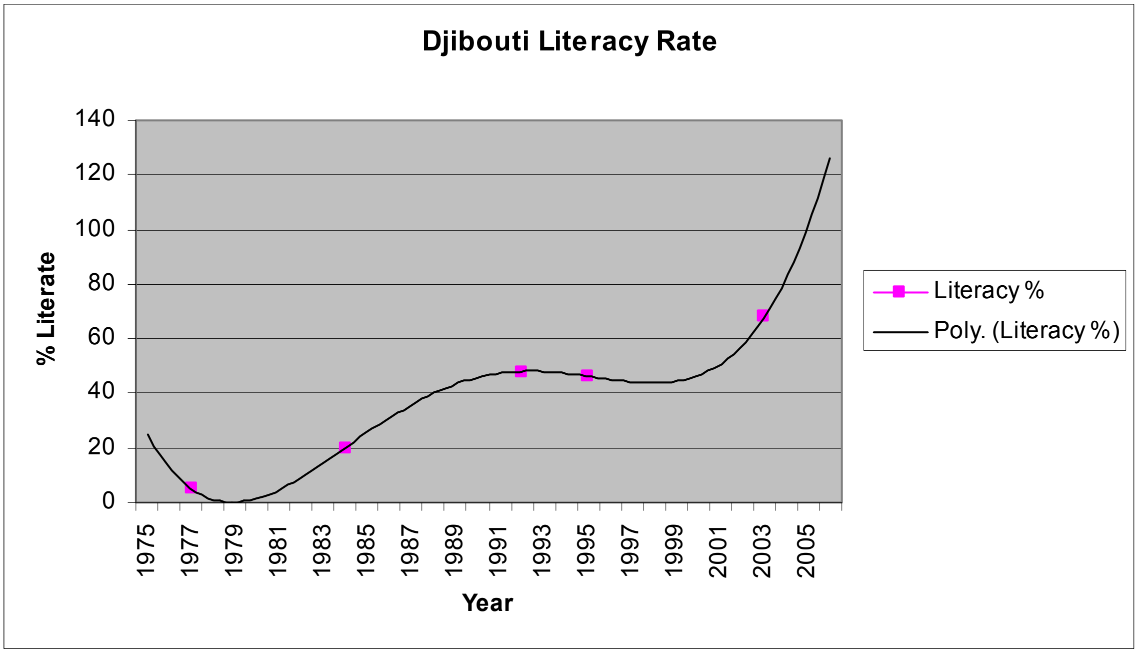


Figure 2-9. Example of Polynomial Interpolation

Cubic Splines. The potential for oscillations in a small portion of a model being able to cause large oscillations and errors over the entire range to be interpolated limits the use of polynomial interpolation. Cubic splines create a separate function over each interval between each set of adjacent points using cubic polynomials, which ensures that the functions are continuously twice differentiable (Burden and Faires, 2001: 142). One of the major advantages of cubic splines is they produce the smoothest possible curve, or in other terms have the least amount of change in slope over the interval defined by the data's extreme points (Schatzman, 2002: 106).

Cubic spline polynomials satisfy the following conditions:

a. The function $S(x_j)$, which defines the line in each sub-interval between real points, is a cubic polynomial

b. The endpoints of each polynomial ($S(x_j)$) yield the correct (known) value of $f(x_j)$.

c. The complete line generated by all the splines must be continuous.

d. The slope (first derivative) of the line where splines meet must be equal for each spline.

e. The change in slope (second derivative) of the line where splines meet must be equal for each spline.

f. One of the following must be true:

1. $S''(x_0) = S''(x_n) = 0$ (free or natural boundary)
2. $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ (clamped boundary) (Burden and Faires, 2001: 143)

Other Types. Trigonometric interpolation, which includes Fast Fourier Transforms, can be used when the time series data has a cyclic or periodic nature. The data used in this study is not seasonal, and when trends are shown they do not exhibit either a cyclic nature or periodicity. Hermite Interpolation can be used in cases where the actual slope of the data is known at each point, as well as the value of the data point itself. This is not know in the data set used here, and is again not applicable. In cases where there is a large amount of noise or variability in the data set splines may not be appropriate. In these cases a least squares fit or a smoothed average may be more appropriate (Burden and Faires, 2001: 164).

2.3.4. Extrapolation, Prediction, and Forecasting Methods

Data sets do not always miss data between known values. In these cases the data must be somehow extrapolated or predicted. In many instances there is information which will assist in the prediction, including a high degree of correlation between two variables or the autocorrelation of data points within the variable itself. Several methods are available to predict data points that are not candidates for interpolation. As with interpolation, extrapolation attempts to give a least error estimate of the data, rather than just a plausible value.

Polynomial Extrapolation. Polynomial extrapolation uses the function describing a dataset to find data beyond the bounds of the known data points. Polynomial interpolation generates a function which is continuous. Generally, the polynomial values very close to the ends, $P(x_0)$ and $P(x_n)$, may provide an estimate. These estimates based on the polynomial usually rapidly diverge due to the exponential nature of the defining polynomial. This can be seen in Figure 2-9, where the last literacy data point is in 2004, and by 2006 the predicted literacy rate is approximately 110%. Therefore, this technique is usually not recommended where the extrapolation will result in erroneous values.

Regression Models. When there exists two sets of data which are highly correlated, but one of them has missing data at its tail(s), a regression model to map the complete data onto the incomplete data may be used. The drawback to this method is it assumes perfect one to one correlation between the two data sets. It biases the data towards a higher overall correlation between the two sets. However, in some cases this may be desirable (Schaefer, 1997: 2).

Suppose there are two data sets for the same variable, but neither covers the entire time span. Since these variables are supposed to measure the same type of data, such as caloric intake per day, but they might be measured slightly differently, mapping one onto the other to create a complete time series does not create correlation bias between two different variables within the data set. However, it may also cause a jump or an induced change point in the data series.

Exponential Weighted Moving Average (EWMA) Model. EWMA is also known as the Single Exponential Smoothing (SES) method. This method uses weights from previous observations via exponential decay to give the most recent events the greatest weight and exponentially less weight as they get older following an decay function with a constant α . Thus, the forecast *F* for *t+1* is represented:

$$F_{t+1} = F_t + \alpha(X_t - F_t)$$

The value α is a parameter between 0 and 1, and can be changed iteratively to find the values that result in the smallest total error, however it is measured. Note that the EWMA estimates always lag the actual values when looking for changes in slope (Makridakis, Wheelwright, and McGee, 1983, 87). Some advantages of this method are that it requires limited data storage, is very easy to code, and not difficult to understand. However, its lag and lack of flexibility reduce its accuracy.

Adaptive Exponential Weighted Moving Average. This is also known as Adaptive Response Rate Exponential Smoothing (ARRES). This algorithm allows the values of α to change over time in order to minimize forecast error. The basic algorithm for ARRSES is:

$$F_{t+1} = \alpha_t X_t + (1 - \alpha_t)F_t$$

where

$$\alpha_{t+1} = \left| \frac{E_t}{M_t} \right|$$

$$E_t = \beta\, e_t + (1 - \beta)E_{t-1}$$

$$M_t = \beta |e_t| + (1 - \beta)M_{t-1}$$

$$e_t = X_t - F_t$$

E_t is smoothed error term and M_t is an absolute error term. β is a parameter between 0 and 1, as is α_t. β can be changed iteratively to find the best value in terms of cumulative error, whereas in this algorithm α_t modifies itself as the program progresses. This method is still relatively simple to code, and not overly expensive computationally. Its flexibility offers some advantage over SES, but can be over reactive with noisy data. (Makridakis, *et al*, 1983, 87).

Autoregressive Moving Average (ARMA) Model. This model combines autoregressive prediction techniques with moving average techniques, and assumes the process is stationary. The basic formulation of the ARMA (1,1) model is:

$$Y_t = \phi_1 Y_{t-1} + \mu^{'} + e_t - \theta_1 e_{t-1}$$

$$\mu^{'} = \mu - \phi_1 \mu$$

where

ϕ_1 is the autoregressive coefficient for a one time unit lag with a value between -1 and 1.

θ_1 is the moving average coefficient for a one time unit lag with a value between -1 and 1.

μ is the mean of all responses.

e_t is error at a particular iteration.

$\phi_1 Y_{t-1}$ is the autoregressive portion of the model.

$e_t - \theta_1 e_{t-1}$ is the moving average portion of the model

It is possible to create lag terms for the ARMA model beyond 0 and 1; however, in practical application it is rare to go further than 2 (Makridakis, *et al*, 1983, 359).

Autoregressive Integrated Moving Average (ARIMA) Model. Similar to the ARMA model except for the allowance of a third term, d, which is the degree of differencing involved. The purpose of this third term is to allow for the modeling of a non-stationary process. ARIMA models are described as ARIMA (*p,d,q*), where p is the order of the autoregressive process, d is the degree of the differential process, and q is the order of the moving average process. The ARMA (1,1) model shown above is equivalent to an ARIMA (1,0,1) model. An ARIMA (0,1,0) model is:

$$Y_t = Y_{t-1} + e_t$$

Again, terms of greater than 2 for *p,d*, and *q* are seldom used (Makridakis, *et al*, 1983, 359). The coding of ARIMA models is difficult due to the interaction of multiple variables that need to be explored to minimize error. However some statistical software packages such as JMP can create optimal ARIMA models based on the input data (JMP, 2006).

2.3.5. Common Imputation Techniques

The imputation methods described in this sub-section were applied only when all other methods of creating time series data have failed or are simply not feasible. This is because most imputation methods ignore autocorrelation or induce variance as a surrogate thus causing a less probable value to be imputed, and they often make the data lose the property of being time series. Neither of these is desirable for the time series data set used in this study. Methods which involve the deletion of missing data are not feasible for this research, since about 50% of the data has to be interpolated, extrapolated, or imputed. Unless stated otherwise, imputation methods described here make the assumption the data is MCAR.

Mean Substitution (Imputation). This method simply uses the mean of the known data, substitutes it for the missing values in that variable response. This method relies heavily on the MCAR assumption, and artificially reduces the variance by a factor of (n^j – 1)/(n – 1), where n is the number of observations, and j is the number of missing observations (Little and Rubin, 1987: 44). While simple to use, this method is not satisfactory for the purposes of this research due to its limitations and the availability of more appropriate methods.

Maximum Likelihood (ML) Estimate. Generates some θ which maximizes the likelihood function L(θ|Y), or equivalently, the loglikelihood l(θ|Y). There can be more than one estimate. In cases where the ML estimate is treated as unique it can be found by differentiating the likelihood function, setting it equal to zero, and solving for θ. The likelihood equation is expressed as:

$$S(\theta \mid Y) \equiv \frac{\partial l(\theta \mid Y)}{\partial \theta} = 0$$

For each component of θ there is an individual equation, resulting a system of equations defined by differentiating l(θ|Y) with respect to all the components of θ. A Multiple linear regression imputation is a special case of a maximum likelihood estimate (Little and Rubin, 1987: 80-83).

Multiple Regression Imputation (Buck's Method). Buck's Method uses linear regression of the complete variable sets to create a model for each variable which contains missing data. The model for each variable is then used to produce an estimate for each missing value. An advantage of this method is that it can be used to forecast data when there are a large number of years to forecast, the usual forecasting techniques are not advisable, and the fitted model is strong. This may produce more "knowledgeable" guesses than some other imputation methods. The method does not work well in cases where there is little covariance between the variables; in such a case Buck's Method produces a model which is very uncertain. It is hazardous to use this to forecast data, however, since it makes an assumption of linearity which may not hold beyond the limits of the observed data. It ignores autocorrelation and can lead to jumps in the transition from real data to imputed data. Buck's Method is also of limited utility when estimating categorical variables (Little and Rubin, 1987: 45-47).

Hot Deck Imputation. A general term for a method which chooses from an estimated distribution of the variable. In most cases this is applied by imputing values from responding units. There are many methods for selecting which value to impute from a responding unit to a non-responding one. Hot deck estimators are unbiased only if the unrealistic assumption is made that probability of response is not related to the true value of the missing data (Little and Rubin, 1987: 60-65).

Nearest Neighbor Hot Deck Imputation (NNI). This is a hot-deck selection method which selects imputed variables based on a computed distance between two observations in a multivariate data set. For each missing data point the algorithm finds a distance between each observation with that variable and the observation missing the data from that observation. The donor observation closest in overall value to the observation missing the data point, referred to as the recipient, then imputes its own value of the missing data to the recipient observation (Little and Rubin, 1987:65-66). One method for finding distance is:

$$D_{1,2} = \frac{\sum_{i=1}^{n} (Y_{1i} - Y_{2i})^2}{m}$$

where

D is the distance between observations one and two

Y_{1i} is the standardized value of ith variable for observation Y_1,

m is the number of values in both observation one and two that both hold real (non-imputed) values

One of the dangers of NNI is its MCAR assumption and it's lack of ability to deal with trends. An example of this would be a country's GDP which has gone up every year, but the data stops in 2000. From 2001-2006 it is likely that the value for 2000 would be imputed, despite a strong belief based on historical trends and other correlated variables that indicate it should still be rising.

Multiple Imputation Data Augmentation (MIDA). MIDA is different from the other methods described so far in that it is a Monte Carlo stochastic simulation rather than a deterministic algorithm for data imputation. Each variable can be fit to a distribution. MIDA uses the distribution associated with each variable to randomly select values for missing data from that distribution. Multiple, different data sets are generated and analyzed individually to determine the sensitivity of the analysis to the imputed data (Rubin and Little, 1987: 255-257). Surprisingly few data sets are required to reduce the standard error of estimates to within a few percentage points of an infinite number of generated data sets. For example, with 50% of the data imputed, five generated data sets would have a standard error only 1.048 times larger than an infinite set of datasets (Rubin, 1987: 114). Implementation of MIDA may be dependent on software availability.

2.3.6. Summary of Missing Data Discussion

Missing data must be overcome to accurately use the statistical techniques proposed. The problem of missing data is common enough in the social and economic fields that there is a significant body of research on the topic. It would be preferable if the set was complete, but several techniques have been discussed to deal with different missing data scenarios. Each problem is different, and every effort was be made to apply appropriate techniques, but where possible and appropriate the research will generally choose these methods in the following order of preference:

1. Interpolation – cubic splines
2. Extrapolation – Adaptive Smoothing or ARIMA
3. Extrapolation – Multiple Regression
4. Imputation – Multiple Imputation Data Augmentation (if available)
5. Imputation – Nearest Neighbor Hot Deck Imputation

These methods are in this order of preference since, in general, when one method is preferred over another it provides a more accurate estimation of the data, particularly when the data is strongly autocorrelated (Horton, *et al*, 2007: 80).

Chapter 3 discusses how each set of variables was imputed individually and why.

2.4. Commonly Used Quantitative Techniques

2.4.1. Ordinary Least Squares (OLS) Regression

OLS is one of the most commonly used and well understood forms of statistical modeling and analysis. In the previous sections regression was referenced as both a method of analysis of data predicting failing states with PITF and as a tool for imputing missing data. It has frequently been used in the fields of political science and econometrics. However, OLS regression is often abused or its limitations and assumptions ignored, all of which can lead to unsupportable conclusions. For the analysis section of this project, OLS regression proved to be a useful tool and framework for a predictive model of undernourishment.

This is not to imply that regression is a perfect tool, or that this research is perfectly suited to be analyzed via regression. Numerous hurdles and potential problems must be overcome; these include multicollinearity, categorical data, covariance, violation of OLS assumptions, and heteroscedasticity. All of these can lead to misinterpreted or misrepresented results that result in an incorrect conclusion. This section examines the basics of OLS and of potential pitfalls of conducting analysis using OLS. Corrective measures will be discussed briefly, but the deeper mechanics and specifics will be left to Chapter 3.

2.4.1.1 Regressors

Regressor variables are also often referred to as predictor variables or independent variables. They represent data that is hypothesized to have a statistical impact on some measurable outcome, known as the response or dependent variable. In this research, many different factors have been hypothesized to be causes or indicators of stability, such as GDP per capita, primary commodity exports as a percentage of GDP, and institutionalized discrimination. In an OLS regression model these factors could be modeled as regressors, since they are suspected of having an effect on some measure of stability as a response.

One common error in the use of regression is the treatment of regressor variables as causes. Even if a model is very accurate at predicting some response and a particular regressor is shown to be significant to that model, it does not mean that the regressor is causing the response (Montgomery, Peck, and Vining, 2006: 5). An example of this would be the data mining done for Osco Drugs in 1992 by Teradata Inc. Teradata's analysts found that the sale of beer and infant diapers both fluctuated throughout the week together, peaking on Fridays between 5 and 7 P.M. (Power, 2002: 1). The sale of beer could be used to very accurately model how many infant diapers would be sold, but the use of infant diapers in no readily understandable way causes the purchase of beer.

Another potential issue with regressor variables is the use of categorical variables. The Minorities at Risk project uses a subjective scale to describe how repressed minorities are in different countries on a year by year basis. It is defined by the following:

> -2 = **Advantaged:** 3 or more checked advantages
> -1 = **Some advantages:** Only 1 or 2 checked advantages
> 0 = **No socially significant differences:** A "socially significant" difference is one that is widely seen, within the minority, and/or the dominant group, as an important distinguishing trait of the group
> 1 = **Slight differentials:** There are socially significant differences between the minority and the dominant group on one or two of the specified qualities. (1 or 2 components checked)
> 2 = **Substantial differentials:** There are socially significant differences with respect to 3 specified qualities
> 3 = **Major differentials:** There are socially significant differences with respect to 4 specified qualities
> 4 = **Extreme differentials:** There are socially significant differences with respect to 5 or 6 specified qualities (Minorities at Risk, 2004: 28-29)

This categorical data poses a challenge to regression. The linear relationships between two numbers does not reflect any sort of physical reality. A score of two does not imply being twice as good as score of one, nor is a score of 1 infinitely better than a score of zero. One potential way around these issues comes from design of experiments by reducing the variable into a single indicator variable, either there is little discrimination (score of -2 to 2) or there is significant discrimination (score of 3 or 4). If the former is true, a score of 0 is given. If the latter, a score of 1 will be substituted. This reduces the problem to one of two level treatment effects, which is discussed Chapter 3 (Montgomery, 2005: 23).

2.4.1.2. Response Variables

The heart of regression is the ability to model the behavior of some response. In OLS there is only one response variable being modeled by one or more regressor variables. If OLS is used, the response variable should be something continuous and linear, and not categorical and subjective. The time it takes to get to work is continuous and linear. The categorized data from Minorities at Risk is not. Other methods exist to model subjective and categorical data, such as discriminant analysis, logistic regression, and poisson regression (Montgomery, Peck, and Vining, 2006: 427).

Different groups have tried to model stability and predict state failure using response variables that were either 0 or 1 for failed or not failed, or on a 1 to 4 scale such as the Center for Army Analyses ACTOR model which used the KOSIMO scale of conflict. Other groups such as the Fund For Peace (FFP) used a nearly continuous scale to assess stability. The problem with the FFP score is it is based on the subjective assessment of subject matter experts as well as an aggregated sum of subjective scores, thus it has the usual problems with a lack of linearity discussed earlier. A single, unified humanitarian crisis score was proposed by Nafziger and Auvinen in a 1997 paper which included number of people killed in battles, infant mortality rate, daily calorie supply per capita, and refugees (Nafziger and Auvinen, 1997: 14-17). For this thesis the research concentrated on stability indicators which represent a point of no return, which are those which effectively preclude intervention by outside entities. War between states, civil war, genocide, politicide, and refugees all represent continuous and quantitative ways to measure a country's instability. War, refugees, and genocide also indicate conditions which the world at large generally lacks the political will to become involved. Battle deaths per capita, genocide and politicide deaths, undernourishment as a proxy for starvation deaths, and refugees per capita each provide a scaled representation of an individual country's plight.

2.4.1.3. OLS Assumptions

There are three basic assumptions of data within an OLS model. First, the relationship between each regressor and the response is at least approximately linear. Secondly, the error terms are normally distributed with a mean of zero and a constant variance σ^2. Third, the error terms are independent random variables which are uncorrelated (Montgomery, Peck, and Vining, 2006: 122). If these assumptions are not met, the problem must be re-defined in some way such that they are, otherwise the validity of the results may be questionable.

The first assumption can often be investigated via visual inspection of the "x" vs. "y" plot of the data for each variable against the response. Figure 2-2 shows a number of graphs which do not show anywhere near linear relationship between the probability of instability and the x variable. Numerous studies show that politically weak democracies have the highest level of instability, whereas strongly autocratic states and full democracies are much less likely to experience instability (Goldstone, *et al*, 2005: 30).

This creates a bell shaped relationship that is not immediately suitable for regression. Percent largest religious group is another good example of a non-linear relationship. However, this problem does not represent an insurmountable obstacle. Numerous options exist for finding invertible transformations of the x data which make it possible to create a relationship which is much closer to linear. They depend on what non-linear relationship between x and y is, and vary from case to case (Montgomery, Peck, and Vining, 2006: 165-166). It is also possible to aggregate or break data down into indicator variables which create treatment effect. There are also more quantitative ways to look for curvature, such as looking at variables for a strong slope but having a poor fit.

Another assumption is having error terms with a mean of zero. The assumption constant variance, also known as homoscedasticity, is important to knowing what level of confidence one has in the model when comparing the residual with the predicted value of the response. When the variance is not constant, then it is called heteroscedastic. Figure 2-10 shows various types of heteroscedasticity and an example of homoscedasticity. When the heteroscedasticity assumption is not met, it means that the accuracy of the prediction, $\hat{y}_i$, varies depending on the value of $\hat{y}_i$. Suppose the variance of the errors increases as the regressor variable increases. This might indicate that the prediction of $\hat{y}_i$ becomes less reliable as $\hat{y}_i$ increases. Unnoticed, this will cause the researcher to overestimate the accuracy of the model in many cases, which is particularly undesirable if one wants an accurate prediction of stability, and a high $\hat{y}_i$ score indicates high instability (Montgomery, Peck, and Vining, 2006: 131). A number of transformations of the response variable, as well as the regressors, are possible to minimize heteroscedasticity.

Among them are Weighted Least Squares (WLS), which reduces the impact a variable with a high variance has on the model. The underlying idea is to make the most reliable data have the greatest effect on the model (Montgomery, Peck, and Vining, 2006: 179-181). Another method for dealing with heteroscedasticity problems in a regression model is the Box-Cox method, which transforms y to correct for non-normality and non-constant variance (Box and Cox, 1964: 211).

The final assumption is the errors are independent and uncorrelated. While autocorrelation is a useful property for extrapolating data, it degrades the accuracy of the model if the errors are autocorrelated in much the same way as heteroscedasticity reduces the utility of the model. Autocorrelation can be a sign that some variable containing important data is missing from the dataset. When autocorrelation in the error terms is present the model is still unbiased, but the regression coefficients are no longer minimum variance. The residual mean square (MS_{res}) is underestimated, giving the undue impression of accuracy (Montgomery, Peck, and Vining, 2006: 475). Finally, confidence intervals based on t and F statistics are no longer applicable (Montgomery, Peck, and Vining, 2006: 478). Residual plots can be used to detect autocorrelation. Durbin and Watson developed a non-parametric test for autocorrelation which tests any specified value of lag (Durbin and Watson, 1951: 159-178). The Durbin and Watson test is included in many software packages. One method that attempts to correct for autocorrelation is Weighted Least Squares (WLS) regression.

2.4.1.4. Multicollinearity

Multicollinearity exists when a column of data, or variable, within the dataset exists as a linear combination, or near linear combination of other variables within the dataset. Mathematically, this can be expressed as if the dataset is represented by a matrix X with a leading column of 1's, if multicollinearity exists then X'X is singular or nearly singular. When multicollinearity exits it causes the variances of the regression coefficients to be high, leads to large covariance values which violate the third assumption of independence between regressors, and the accuracy and utility of the model to become questionable (Montgomery, Peck, and Vining, 2006: 109). Multicollinearity always exists in raw economic data to some degree. It is often a question of the degree of multicollinearity, its effects, and if these effects are acceptable.

There are both tests and remedial measures to deal with issues of multicollinearity. One is a simple screening process where the correlation between each variable is computed; when two variables are highly correlated, the one with more imputed data is removed from the set. A record of which variable is serving as a proxy for another is kept. This improves (reduces) the variance of the regression coefficients, and since the two variables were very similar, they both had approximately the same linear effect on the model. Unless their correlation was exactly one, however, some information within the deleted variable will be lost from the model. Another problem with this method is that it precludes tests for interaction terms involving the deleted variable. The threshold for deletion based on correlation is left to the modeler.

One simple test for multicollinearity is plotting variables on both the x and y axis and looking for a linear relationship. Others involve using the diagonal values of the covariance matrix C, where $C = (X'X)^{-1}$, as Variance Inflation Factors, or using the eigenvalues of X'X to identify variables with strong multicollinearity. In addition to deleting highly correlated variables, other solutions to multicollinearity include collecting more data and redefining the model by modifying the problematic data in some non-linear way that preserves the data contained in the variables, such as Factor Analysis (FA) score or Principal Component Analysis (PCA) scores (Montgomery, Peck, and Vining, 2006: 342).

2.4.1.5. Interpreting the Model

One of the first questions that a regression model serves to answer is which variables are important to the model. Another is how strong the relationships between the regression coefficients and the outcome are. The regression coefficients' (β) magnitude for each variable is not necessarily an indicator of significance. A very large value might indicate a large response variable and a very small regressor. A value near zero might only be an indication that the relationship between the regressor and the response is non-linear or mis-specified (Montgomery, Peck, and Vining, 2006: 56). A brief synopsis of math behind OLS regression is needed to specify tests for the significance of variables in a model.

In general, following the development given by Montgomery, *et al*. an OLS regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon$$

(Montgomery, Peck, and Vining, 2006: 63)

where

y is the response value

x_1, x_2, … are values from each variable in the model for a particular observation

β_1, $\beta_{2,...}$ are the regression coefficients associated with each variable $x_1, x_{2...}$

ε is the error term

In matrix notation

$$y = X\beta + \varepsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(Montgomery, Peck, and Vining, 2006: 68)

n is the number of observations

k is the number of variables

The model is created by finding the estimated coefficients of each variable, $\overset{\wedge}{\beta}$ via the equation $\overset{\wedge}{\beta} = (X^{'}X)^{-1}X^{'}y$

Estimated values of the response are generated using $\overset{\wedge}{\beta}$, where $\hat{y}$ is the estimate of the responses found with the equation $\hat{y} = X\overset{\wedge}{\beta}$. The sum of squares residual, which measures the difference between the predicted values generated by $\hat{y} = X\overset{\wedge}{\beta}$ and the actual response values, is expressed as SS$_{\text{Res}}$ = $y'y - \overset{\wedge}{\beta}'X'y$. The regression Sum of Squares (SS$_{\text{R}}$) represents the amount of variance explained by the model, and is expressed as:

$$SS_R = \overset{\wedge}{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

(Montgomery, Peck, and Vining, 2006: 63-82)

The total Sum of Squares can be expressed as:

$$SS_T = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

(Montgomery, Peck, and Vining, 2006: 81)

The relationship between SS_{Res}, SS_R, and SS_T is therefore $SS_T = SS_R + SS_{Res}$.

The Residual Mean Square Error (MS_{Res}) is an unbiased estimator of the variance of the model, $\hat{\sigma}^2$.

$$MS_{\mathrm{Re}s} = \frac{SS_{\mathrm{Re}s}}{n-k-1}$$

Similarly, the Regression Mean Square Error (MS_R) is defined as:

$$MS_R = \frac{SS_R}{k}$$

(Montgomery, Peck, and Vining, 2006: 76)

All of this leads to two tests designed to examine the significance of the $\hat{\beta}$ coefficients. The first is the F-test. Since SS_R/σ^2 follows a χ_k^2 distribution, the significance of all the coefficients can be tested together with an F_0 statistic defined by $F_0 = MS_R/MS_{Res}$. When $F_0 > F_{\alpha,k,n-k-1}$, where α is the chosen level of significance, the null hypothesis (H_0) that $\beta_1 = \beta_2 = ... = \beta_k = 0$ can be rejected. This test is not sufficient however, to judge each variable individually for significance (Montgomery, Peck, and Vining, 2006: 85

The more useful of the two tests is the marginal *t* test. Adding more regressors variables to the model will always increase the SS_R and decrease SS_{Res}; however, the trade off is the variance of the fitted value. Adding an insignificant regressor variable to the model may also increase MS_{Res} and thereby the usefulness of the model. Thus a test statistic for each variable is needed to test the null hypothesis $H_o : \beta_j = 0$, and the alternative $H_1 : \beta_j \neq 0$. The test statistic is:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

where C_{jj} is jth the diagonal element of $(X'X)^{-1}$. The null is rejected if $|t_0| > t_{\alpha/2,n-k-1}$ (Montgomery, Peck, and Vining, 2006: 84). The result is a test of significance for regressor j given the other regressors in the model. This test is only a true test of significance if OLS assumptions have been met.

Another aspect of a regression model is how well the response variable is modeled. There are several techniques and statistics in regression designed to address this question. The three statistics examined here are R^2, R^2_{Adj}, and R^2_{pred} all of which range in value from 0 to1. R^2 is calculated as SS_R/SS_T. This statistic simply tells how much of the variance in the model is accounted for by the regressors. R^2 as a measure of model effectiveness has drawbacks. No matter how many regressors are added to the model they will always increase R^2. Use of this statistic can lead to over fitting the model, and does not encourage a parsimonious use of regressors. Thus, use of R^2 leaves out important information regarding the validity and utility of a model.

Another commonly used statistic to describe the adequacy of a model is adjusted R^2. This is defined as:

$$R^2_{Adj} = 1 - \frac{SS_{\mathrm{Res}}/(n-k-1)}{SS_T/(n-1)}$$

The advantage of this statistic is it will only increase when a variable is added to the model which reduces the residual mean square. R^2_{Adj} is therefore useful in choosing variables to include in the model as well as helping to reduce its complexity (Montgomery, Peck, and Vining, 2006: 83-84).

The final R^2 statistic discussed here is R^2_{pred}. This statistic is based on the PRESS statistic first suggested by Allen in 1971 (Allen, 1971: 16). It is defined as:

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

where:

h_{ii} is the value in the i^th^ row and column of the n x n hat matrix defined by $X(X'X)^{-1}X$

e_i is the error for each dependent variable in the model

PRESS is generally used as a measure of how well a model will perform at predicting new data. PRESS is then used to create R^2_{pred} defined by:

$$R^2_{pred} = 1 - \frac{PRESS}{SS_T}$$

The result indicates the percentage of variability expected to be explained by the model for new observations, whereas R^2 describes only the amount of variance described by, and used in building, the model (Montgomery, Peck, and Vining, 2006: 141-142) .

2.4.2. Principal Component Analysis (PCA)

Principal component analysis is a multivariate analysis technique that is useful in regression, reducing variable sets, normalizing data, and explaining the variance-covariance structure of a set of variables using a linear combination of orthogonal vectors. The method can also aid in interpreting the data (Johnson and Wichern, 2002:

427). One of the goals of data reduction is to summarize the data in as few variables as possible while minimizing the loss of information. Given a dataset X, PCA follows the following general algorithm:

1. X_{cr} = COR (X) where X_{cr} is the correlation matrix of X.

2. Find the eigenvectors and eigenvalues of $X_{cr.}$

3. Eliminate the columns of eigenvectors associated with eigenvalues less than one. Component scores with eigenvectors less than one represent inputs with effects less than that of noise within the model, and need not be considered. As a result, the sparse model contains less variance.

4. X_c = X * Reduced set of eigenvectors

The resultant component score matrix X_c may be better suited for regression in place of the original dataset X due to the orthogonality of its variables, which helps satisfy the assumption of independent variables. Each column within X_c is referred to as a component and the individual exemploars in the components are called the scores. These scores, while subject to interpretation, can be treated as any other set of data points describing each observation in the data set. Indeed, the number of scores will be equal to the number of observations in the raw data

The transformation matrix which creates the PCA scores is referred to as the loadings matrix; it shows the contribution of each variable to each principal component. One drawback to PCA is that it can difficult to tell what the meaning of each component is based on the loadings. Thus the labeling of components is somewhat subjective. The loading issue detracts from the purpose of creating a model which is easily interpretable and shows direct relationships between individual variables and stability (Mardia, Kent, and Bibby, 1979: 209-217). As a tool for creating a more powerful predictive model, it has many useful qualities.

2.4.3. Cannonical Correlation

The idea behind canonical correlation is simple. Given two sets of multivariate data X_1 and X_2, find $\eta = a' X_1$ and $\phi = b' X_2$ such that the correlation between η and ϕ is maximized (Mardia, Kent, and Bibby, 1979: 281). This represents finding the correlation between two sets of linear combinations of variables, and is related to PCA. However, unlike PCA where the relationships between the original variables are lost, canonical correlation attempts to find the maximal relationship between different sets of data expressed as two similarly sized matrices of scores. If the data is standardized prior to performing canonical correlation, then the *a* and *b* transformation (loading) matrices shows how each variable is loaded relative to each other on their respective canonical correlation scores contained in the η and ϕ matrices. These loadings will be approximately normal if the data has been standardized, with a mean of 0 and a variance of 1. They can be interpreted in much the same way as PCA loadings, but also suffer from the same interpretability issues (Johnson and Wichern, 2002: 427).

2.4.4. Discriminant Analysis

DA is a multivariate classification technique which can build a model which will separate observations into mutually exclusive and exhaustive *a priori* groups based on training data. This is accomplished by rotating the planes of observation of the groups in order to maximize the between group variance and minimize the intra-group variance. Each object is assigned a score based on its distance from the centroid of each group, and this distance is transformed via a projection to create a probability that the observation belongs to a particular group. The advantage of DA is it will create a model which produces the smallest number of misclassifications of the training data possible using any linear combination. Another way of thinking of this is finding the center of groups of data points, and then measuring how far it is from each center to a new observation. Whichever distance is smallest represents the group which the new observation belongs to. Figure 2-10 gives a basic visual depiction of the concept (Dillon, 1984: 362-363).

An advantage of DA is that it is simple, well understood, and supported. However, given non-normal data, the types of errors produced by DA will be biased, and some of the standard techniques for analyzing the impact of the entering variables will not be valid.

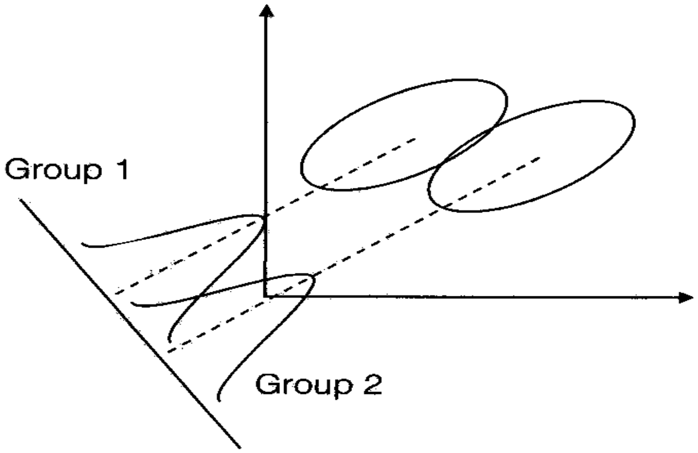


Figure 2-10. Graphical Depiction of Discriminant Analysis in 2-D Space
(Lattin, et al, 2003: 431)

2.4.5 Logistic Regression

Logistic regression is another technique for classifying observations into discrete, mutually exclusive, and exhaustive *a priori* groups based on training data. Like DA, logit regression utilizes variance to distinguish between *a priori* groupings. However, unlike DA, it does not utilize a linear model. The response function of a logistic regression is an S-curve and has a non-linear model:

$$E(y) = \frac{1}{1+\exp(-x'\beta)}$$

(Montgomery, *et al*, 2006: 429)

Given the probability of an observation belonging to any particular group being between 0 and 1 inclusive logistic regression maintains these boundaries. If the equation above is to be linearized, it must be transformed using the following:

$$\eta = x'\beta$$

$$\eta = \ln\frac{\pi_i}{1-\pi_i}$$

Where π is the probability an observation belongs to group *i* and $\eta = \ln\frac{\pi_i}{1-\pi_i}$ is referred to as the logit transformation of the probability of π.

The value of β is also determined using non-linear mathematics via a Maximum Likelihood Estimator (MLE) function. The output of a logistic regression is a series of probabilities describing the chance each observation has of belonging to each *a priori* group (Montgomery, et al, 2006: 427-428). This technique was found by the PITF to have stronger predictive abilities than more complicated statistical methods such as neural networks (Goldstone, *et al*, 2005: 11).

2.4.6. Summary of Techniques

Numerous techniques have been used in the past to model stability. Most have attempted to make a categorical prediction that places states into categories such as core, gap, stable, or unstable. In binary outcome cases logistic regression has been attempted.

In cases with more than two categories Discriminant Analysis and Factor Analysis have been used. Some models, such as the ACTOR, model have blended multiple mathematical techniques. Others have tried simulations. Less mathematically oriented models have been created by agencies such as the FfP and USAid to allow subject matter experts within their agencies to make assessments of conditions for stability. The purpose of this research is to attempt to create a more refined model which predicts the expected level of instability factors three or four years hence in the Horn of Africa region. Regression is well suited to making estimates of continuous data. This project was not wed to any one method of regression, but explored different options for building a model that also utilizes related multivariate analysis techniques.

2.5 Chapter Summary

Numerous attempts at modeling stability have been made. Some have been predictive, and some have been descriptive. Forecasting models are more interesting to economic, military, and humanitarian organizations. The missing data in the data set is imputed using interpolation, extrapolation, and multiple imputation. This research utilizes the tools such as canonical correlation, PCA, DA, and logistic regression, to create accurate predictive models of instability indicators using the constructed data set. These models are examined to determine which variables significantly contribute. Chapter 3 describes the variables in the data set, missing data imputation methods used on the data set, and the mathematical techniques used to build the forecasting models are described in greater detail. Chapter 4 describes the models developed using the techniques described in Chapters 2 and 3.

3 Methodology

3.1. Introduction

The primary purpose of this study is to develop a accurate longer term prediction models of instability in the Horn of Africa, with a secondary goal of identifying variables that contributed most significantly to the models. This chapter presents an overview of the methodologies used in the study. This includes the variables selected for use in this study, the methods used to analyze them, and how the model was built. The chapter begins with a discussion of the data search and the variables that were ultimately used in the data set. Next it discusses the interpolation, extrapolation, and multiple imputations used to deal with missing data. The chapter concludes with a discussion of the methods used to build competing models, how the models are compared, and how the models can be compared with previous efforts. Figure 3-1 shows the general flow of the process used by this study. Chapter 3 describes in detail the steps in data collection, and the methods and techniques used the Analysis phase. It also gives a brief description of the models developed. Chapter 4 describes the development of the models, and the results of the analysis phase.

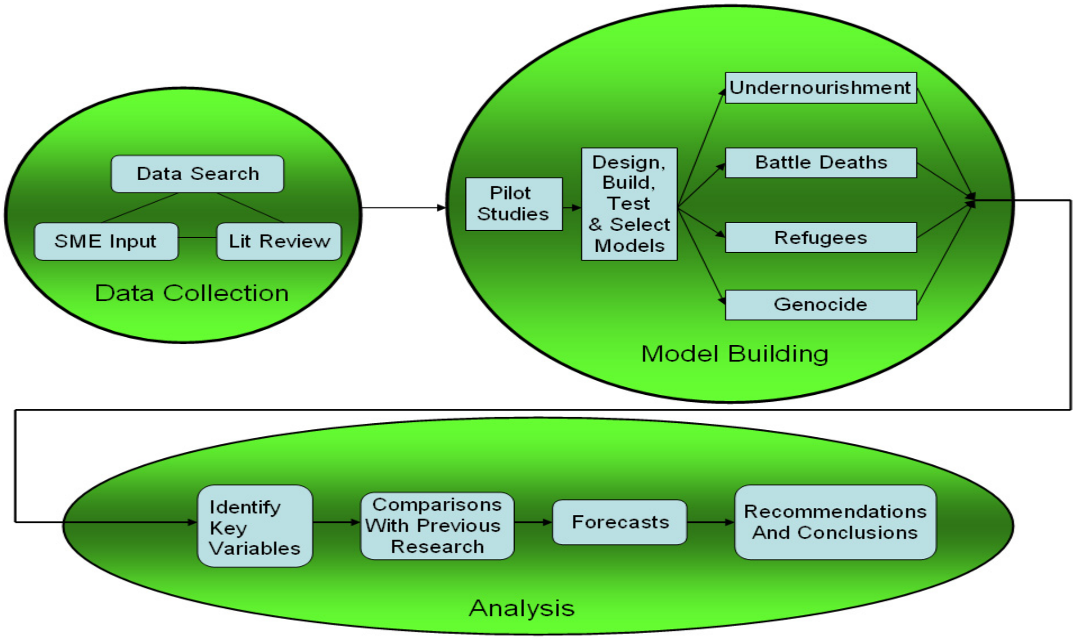


Figure 3-1. Flow of Research

3.2 Data

This section discusses which variables were gathered, which ones could not be located or lacked sufficient data to provide value to the model, or where too much data would have to be imputed. The sources of the data, and the definitions of each variable are provided in the Appendices. The methods used to build a complete dataset from the initially gathered partial set are also discussed.

3.2.1 Initial Dataset

The selection of data started with a review of literature on the subject of predicting failing states, as well as the classification of failing states in previous research efforts. Many of these prior studies were discussed in the previous chapter; a complete listing of the collected variables and their sources is shown in Appendix A. The Appendix shows the data, the source, and which authors or subject matter experts suggested its use. Others, such as change in caloric intake, and interactions between arable land scarcity, reliance on agriculture, and water scarcity have been hypothesized by the author to be potentially significant. Appendix A gives a list of the variables collected. Appendix B gives a definition of all the variables not previously defined in Chapter 2. Finally, Appedix C lists the derived variables.

3.2.2. Variable Selection

The initial data collection was made to capture data for each country from 1975-2006 for each variable cited by the SMEs as crucial to any prediction model of instability indicators. This effort was not entirely successful for availability reasons. For example, Eritrea only seceded from Ethiopia in 1991, making data on the region very sparse prior to the secession. North and South Yemen reunified in 1990, making data prior to that date also difficult to obtain. While approximations might be attempted, they would add more uncertainty if an accurate basis is not established. Other variables identified as contributors were not available at all for many countries, making imputation very imprecise due to the lack of knowledge of what effect an individual country's "treatment" had on a variable. In other cases, the data came from too many sources, was too sparse, or was collected in too many ways to easily build a mapping function.

Clearly, the list of variables in Appendix A does not encompass every measure suggested by SMEs in Chapter 2. Some variables were not included due to paucity of data. The data was simply unavailable, or too sparse to be imputed without the noise and variance being more than the actual data contained in it. In particular, the crime data suggested by Capt. Robbins as a strong measure of internal stability and governmental ability to provide security was unavailable (Robbins, 2007: 37).   Capt. Nysether had gathered crime data for his global study, but when examining the seven countries in the

HoA there were only 9 data points available out of a possible 618 (Yemen's total crimes, drug crimes, and attempted murders for 1998, 1999, and 2000).

The GINI coefficient is a measure of disparities within a society, whether the inequality is measured by income, land, or another continuous variable. High GINI scores are considered to be indicative of a very uneven distribution of assets. Collier and Hoeffler postulate this inequality of wealth to be the underlying cause of civil war (Collier and Hoeffler, 1998: 7), (Collier and Hoeffler, 2004: 5). Unfortunately, the GINI coefficient was incomplete and inconsistent for the countries of interest; when it was available it had been sampled by different agencies using different methodologies resulting in different answers. The result was 7 data points for 7 countries, measured 3 different ways. As this inconsistency was deemed not to be acceptable, the Minorities at Risk political discrimination and economic discrimination scores were used as a proxy for the GINI coefficient. While the Minorities at Risk scores were treated as a binary categorical variable, the amount and reliability of the available data was judged to be an acceptable trade off despite the loss of resolution.

Forty-five independent variables were included in the research. A complete list is shown in Figure 3.1. Detailed descriptions of the source of the data, and how calculated dated was created are included in the Appendices. Section 3.2.3 discusses special cases of how the data was handled.

Agriculture as % GDP	Literacy
Aid as % GNI	Military as % GDP
Aid per Capita	Missing Data
Anarchy	Trade Openness
Arable Land Per Capita	Partial Autocracy
Bad Neighbors	Partial Democracy w/Factionalism
Youth Bulge	Partial Democracy w/o Factionalism
Change in Calories	Paved Roads as a % of Total
Change in Infant Mortality Rate	Principal Commodity as a % of Exports
Country	Political Discrimination
Durability	Population Density
Economic Discrimination	Population
Ethnic Fractionalization	Religious Fractionalization
Forested Land	Telephone and Cell subscribers / 100
Full Autocracy	Trade Ratio
GDP Growth	Transition Governments
GDP Per Capita in 1998 USD	Urban Population as a % of Total Roads
Gender Parity	Water / Agriculture Interaction
Infant Mortality Rate	Water per Capita
Km Road Per Capita	Water/ Agriculture/ Land Interaction
Land Stress	Year
Life Expectancy	Years since last conflict
Linguistic Fractionalization	

Table 3-1: Independent Variables

Table 3-1 shows the independent variables collected for this study. Appendices A, B, and C provide sources, interaction terms, and definitions of each. The dependent variables for this study's models are battle deaths per capita, genocide and politicide deaths, refugees per capita, and percent of the population designated as "Undernourished" by the UN FAO Undernourishment is used in place of for starvation deaths, with the expectation that the two are highly correlated. Undernourishment may, however, act as a precursor of starvation. The quantitative dependent variable "refugees per capita" was used in place of the PITF's more subjective dependent variable of "adverse regime change" (Goldstone, *et al*, 2005: 7).

3.2.3. Special Data Handling

Initially, the plan was to use data from 1975 to 2006. The dates were chosen based on the available data becoming very sparse prior to 1975. Data for 2007 generally has not been collected or compiled at the time of this writing, and data collection for this study took place in CY 2007. While a longer time series allowed for more historical incidents of instability and periods of stability, when using time series analysis tools such as ARIMA and the Amelia II multiple imputation software a larger period is more likely to cover changes in the structure of the data and the model. Complete data for the period did not prove to be entirely feasible in several cases, however.

Djibouti was a French protectorate until June 27th 1977. As such, very little economic data was found for years prior to 1977. The data used in this project excluded data on Djibouti prior to 1977 (CIA World Fact Book, 2007). Eritrea's data posed a similar problem. Eritrea did not win its 30 year war of independence against Ethiopia until 1991. Its government was not fully in place until 1993 when they gained full international recognition. Data prior to 1991 on Eritrea was not considered. The exception to these cases is data before 1977 in Djibouti and 1991 in Eritrea where 1975 and 1990 data points allowed interpolation to be used to obtain better estimates of missing data in the remaining years considered. Because of all of these factors, data on Eritrea used in this study reflects the years 1991 though 2006.

Yemen presented another problem. This country was included in the study due to its extremely strong economic ties to other countries in the Horn of Africa, including Sudan, Djibouti, and Somalia. It is also the poorest country on the Arabian Peninsula. Prior to 1990 Yemen was divided into North and South Yemen. Additionally, Yemen is in the CENTCOM area of responsibility as well as the other HoA nations. The somewhat inappropriately named North Yemen had more than 2/3rds of the total population and controlled of most of the country's urban areas and natural resources. As a result, North Yemen was significantly richer than the Marxist South Yemen. However, the two reunified relatively peacefully in 1990. Since North Yemen held the vast majority of wealth and population, the resulting post-1990 Yemeni government resembled the North Yemen government far more than the South Yemen government, much the same way the resultant re-unified Germany resembled West Germany much more than East Germany (CIA World Fact Book, 2007). The problem with the dataset is due to the lack of data prior to 1990 for Yemen as a whole, North Yemen, or South Yemen. Economic data available does not discriminate between the two countries, and their social and political data were very similar. Additionally, North Yemen controlled the coast line from which commerce with the Horn of Africa is conducted. Thus, in the cases of government durability, years since last war, executive competition, government restrictions on political competition, and the binary variables describing type of government, data from North Yemen was used from 1975 through 1989.

Another unique aspect of Yemen is its fractionalization scores. Yemen is almost completely homogenous in terms of religion, ethnicity, and language. In fact, there is no sub-group large enough that even if suppressed, it would not be registered as a systematic persecution by the Minorities at Risk (MAR) Project. Thus, MAR does not collect statistics on Yemen. For this study's dataset Yemen's political and economic discrimination scores were set to zero (no discrimination), since there are essentially no ethnic, linguistic, or religious minorities to bar from political or economic life there (e-mail with Dr. Pate, 11/06/07).

In a few cases existing data was deliberately deleted due to indications it may have been falsified. Notably, some government supplied figures on literacy were significantly different from the UN estimates and the CIA World Fact Book Estimates. The two cases that were deleted from the data were Eritrea's declared 58.6% literacy rate in 2005 and Somalia's 1989 self estimate of 60%. The previous data points were 25% in 2000 and 10% in 1984 respectively. Alternate data points for the years in question were not found via other sources. Such vast increases seem unlikely, and do not reasonably match data from the other sources used (UN Common Database, 2007), (CIA World Fact Book, 2007). The illiteracy rates for Eritrea in 2005 and Somalia in 1989 were instead interpolated using surrounding data points.

The Upsala Conflict Database was used to find, via inspection, both the "bad neighbors" data as well as the independent variable "years since last conflict". The database only covers the years from 1945 through 2006. Somalia and Djibouti did not have any listed conflicts from 1945 through 1975 (or 1977 in Djibouti's case). However, Djibouti and Somalia were both territories of countries involved in World War II (France, UK, Italy). Thus, for the "years since last conflict" variable, 1945 was treated as the last year of conflict for both Djibouti and Somalia, and calculated forward accordingly.

In some cases there were no data points available for the most recent years. Consultation with those responsible for maintaining and updating the Minorities at Risk database, as well as the POLITY IV data was conducted via e-mail to verify the scores on countries that have not been released as of yet to the public. Dr. Amy Pate confirmed that the political discrimination data on the countries in this study for 2004-2006 has not changed (Dr. Pate correspondence, 11/06/07), (Dr. Marshall correspondence, 09/18/07). . Dr. Monty Marshall with the POLITY IV project indicated that there are no current changes to the 2005-2006 variables of interest from executive recruitment and government restrictions on political competition their dataset from the 2004 value.

3.2.4. Categorical Variables

In Chapter 2 the problem of categorical variables was discussed; the relationship between variables is not linear, nor does an expected, imputed, extrapolated, or imputed fractional score have any meaning. However, in the dataset for this study there are several categorical variables, including the MAR and POLITY IV scores. Montgomery suggests that a way of dealing with this is to use Design Of Experiments (DOE) methodologies and break each variable down into a two level binary treatments. However, unlike truly designed experiments, the dataset is historical and not subject to experimentation, randomization, replication, or choosing the settings of the variables (Montgomery, 2005: 13-14).

The method of breaking down the executive recruitment scores and political competition scores was suggested by PITF in their PITF V report, and laid out explicitly in correspondence with Dr. Jay Ulfelder. Governments were separated into several categories in PITF V: Full Autocracy, Partial Autocracy, Partial Democracy with

Factionalism, Partial Democracy with Factionalism, and Full Democracy. In this study, two more government types were used: Transition Government when the POLITY IV score was -88, and Anarchy when the score was -77. The PITF V project did not consider data on countries which were in transition or anarchy; however, this study does. There have been no full democracies in the HoA region during the study period, and thus the category was omitted as a category. A complete description of how the government types are scored and defined is provided in Appendix C (Dr. Ulfelder correspondence, 10/10/07) .

For this study the MAR data selected the worst score for any listed group in a particular year for a given country, and recorded it. This was done for both economic and political discrimination. After each country was given a single score between 0 and 4 for the two variables, the scores were converted to a zero or a one. When the POLDIS or ECDIS score was greater than or equal to three, the score became a one indicating systematic discrimination took place. When the score is 2 or less, it is changed to a zero indicating no systematic discrimination took place. While this decision was subjective, it was based on the definitions provided by MAR, and seems to form a clear distinction between the high and low treatments.

3.2.5. Interpolation

After gathering the variable set 19.3% of the data was missing. The variables listed in Appendix B were excluded from this calculation because there are calculated based on other variables, therefore their "missingness" is completely dependent on originally collected variables. The 19.3% figure also represents the data set after categorical data and assumptions as described in Section 3.2.4 have been made and entered. Table 3-2 shows the variables used to determine the missing data percentage.

Agriculture as % GDP	Linguistic Fractionalization
Aid as % GNI	Literacy
Aid per Cap	Malnutrition
Anarchy (-77)	Military as % GDP
Bad Neighbors	Missing Data
Battle Deaths as % of Pop	Trade Openess
Youth Bulge	Partial Autocracy
Calories Per Day Per Capita	Partial Democracy w/Factionalism
Change in Calories	Partial Democracy w/o Factionalism
Change in IMR	Principal Commodity Exports as % GDP
Country	Percent of Roads that are Paved
Durability	Political Discrimination
Economic Discrimination	Population
Education as % GNI	Population Density
Ethnic Fractionalization	Refugees as % Pop
Forested Land	Religious Fractionalization
Full Autocracy	Telephone Subscribers per 100 Population
GDP Growth	Trade Ratio
GDP98	Transition (-88)
Gen/Poli per Capita	Urban Population Percentage
Gender Parity	Water Per Capita
Infant Mortality Rate	Year
Km Roads	Years since last conflict
Life Expectancy	

Table 3-2. Variables Used to Calculate Percentage of Data Missing

The first step used in replacing missing data in this study was to impute data via interpolation. Cubic Splines were used almost exclusively to interpolate the data, based on the smoothness of the curve they produce, which is appropriate to the data set in this study. The exceptions to this were the limited number of cases where using cubic splines resulted in impossible values, or there were only two data points. Iterpolation data is shown in Appendix D. Chapter 2 briefly described the method of cubic splines, it is described in further detail here in the next sub-section.

The mathematical development of cubic splines below is based on the development given in Burden and Faires' text (Burden and Faires, 2001, 142-146). Following their development; given that each piecewise section between points x_j and x_{j+1} is expressed as a third degree polynomial function of the form:

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

(Burden and Faires, 2001, 142-146)

for each j = 1,2,….*n-1*.

As expressed in Chapter 2, when condition c. from 2.3.3, which specifies the entire function must be continuous, is applied to the equation the function becomes:

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3$$

where

$$h_j = x_{j+1} - x_j$$

Similarly, condition d, the slope (first derivative) of the line where splines meet must be equal for each spline, thus defines $b_n = S'(x_n)$ resulting in the equation:

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^3$$

Therefore, $c_n = S''(x_n)/2$, and condition e. results in the equation:

$$c_{j+1} = c_j + 3d_j h_j$$

For each j = 1,2,….,*n-1*.

Given the system of equations above which satisfy conditions a through f for cubic splines, variables a_j, a_{j+1}, b_j, b_{j+1}, d_j and d_{j+1} can all be replaced by substituting variable *c*. Solving for d_j in the equation $c_{j+1} = c_j + 3d_j h_j$ yields:

$$d_j = \frac{(c_{j+1} - c_j)}{3h_j}$$

and substituting it into the previous equations for a_{j+1} and b_{j+1} results in:

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1})$$

and

$$b_{j+1} = b_j + h_j(c_j + c_{j+1})$$

Substituting b_{j+1}'s equation into the equation for a_{j+1} yields

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1})$$

If the index *j* is reduced by one, the system of equations reduces to:

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1})$$

For each $j = 1,2,\ldots,n-1$ (Burden and Faires, 2001, 144).

This system of equations is now solvable because the vector of variables c_j is the only set of unknowns, since the value of each a_j is already known as the value of $f(x_j)$. Solving for c is now simply a matter of solving the vector equation Ax = b where

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}$$

and b and x are the vectors

$$b = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \quad x = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}$$

The linear system will have a unique solution to *x* (Burden and Faires, 2001, 146).

There are two general types of cubic splines: clamped and free (or natural). Clamped splines provide better estimates of the defining functions, but require that the slope of the function be known at the endpoints. For the data in this study, these values were not known and therefore free boundary cubic splines were used (Burden and Faires, 2001: 143). All cubic splines in this study were calculated using the MATLAB 2007b built in *spline* (X,Y, xx) function, where X is the vector of values for a variable for one country over the years Y. The pre-defined vector xx indicates what values to interpolate the values X over.

3.2.6. Extrapolation

Following the interpolation step, 14.6% of the data was still missing. This study used various types of ARMA and ARIMA models to extrapolate that data which was amenable to the process. This was done for several reasons. In every case, an ARIMA model could be found that provided a better R_{adj}^2 than any of the exponential models. No readily available model was found for ARRES. This is not unexpected. EWMA smoothing is equivalent to an ARIMA (0, 1, 1) model, and the other models can be approximated with different ARIMA models (Box and Jenkins, 1976: 106). Thus, the flexibility and availability of supporting software made ARIMA the overall model for extrapolation in this study.

Appendix Y provides a detailed discussion and description of ARIMA (*p,d,q*) models and how they were used to extrapolate data into this study.

3.2.7. Multiple Imputation

Missing data is a common problem in longitudinal social science research. The literature review found references to missing data imputation in medicine, politics, econometrics, political science, and statistics (Horton and Kleinman, 2007: 1). In the field of stability prediction surprisingly little discussion of how projects such at PITF, CAA, and Collier and Hoeffler addressed this issue. Nysether addressed the issue via means which were user implementable, nearest neighbor hot deck imputation. One of his follow on suggestions was to attempt this type of research using multiple imputation (Nysether, 2007, 120). In his case, the primary issues for not pursuing the option were time and software availability constraints.

Several different multiple imputation packages were examined for this research. The SAS function PROC MI is available in SAS 8.1b and later but require proper site licenses. Another popular package is the Multiple Imputation by Chained Equations (MICE) software. However, there are known errors in the program, and it deals poorly with issues of multicollinearity (Horton and Kleinman, 2006, 87). Another widely used freeware program is Dr. Joseph P. Schaefer's NORM program. This program was found to be relatively inflexible for this study. In addition, problems importing data were discovered. NORM also dealt poorly with multicollinearity, and does not directly support general purpose regression. After a review of these packages, the freeware

Amelia II program written by Dr. Gary King of Harvard and Dr. James Honaker of UCLA was used (King, *et al*, 2001: 1).

There were several reasons for selecting this software. First, it proved to be very robust; able to handle matrices of a large size and varying degrees of multicollinearity (which caused some of the other software to crash). It was simple and intuitive to use, and as a software package dedicated to multiple imputation, the users manual proved straightforward. The basic “engine” of AMELIA II has been available since 2001, and has been continually refined by its developers. The algorithm driving the multiple imputation software, and the software itself have been extensively examined in peer reviewed articles (Horton and Kleinman, 2007: 84), (King, Honaker, Joseph, and Scheve, 2001:49-69), (Horton and Lipsitz, 2001: 248-249). Finally, Amelia II incorporates Time Series Cross Sectional (TSCS) data handling algorithms that were designed with variables broken down by country and time period specifically in mind (Honaker and King, 2007: 1).

The basic algorithm underlying the Amelia II software was outlined by King, Honaker, *et al* in the 2001 article “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation.” The algorithms all make the assumption variables are NID with a mean vector μ and a covariance matrix $\sum$, however, in practice this is not necessary to obtain an imputed variable with minimized MSE (King, *et al*, 2001 53-54). This specification of multivariate normal distributions implies the variable can be imputed linearly. There are two general types of multiple imputation algorithms: Imputation-Posterior (IP) and Expectation-Maximization (EM) (King and Honaker, 2007: 8). The Amelia II software uses a third generation descendant of the original EM algorithm. EM works much like IP, except instead of random draws from the estimated distribution of the variable, it replaces missing data with its predicted value. The predicted value is the maximum posterior estimate obtained via OLS to obtain an estimate of β for each variable, then an estimate of the missing data is obtained using the regression equation. The vector of coefficients β is then recalculated using the estimated data. Usually, the starting values of the missing data are simply the mean value of the vector of variables. This process is repeated until β converges. There are problems with this algorithm, however. EM is not true MI, since EM lacks a mechanism for inducing randomness or creating multiple sets. Because it ignores estimation uncertainty, the standard errors are generally biased downward (King, *et al*, 2001: 55) .

The descendant algorithm EMs (EM with sampling) is based on EM. EMs enhances EM by running EM and using the maximum posterior estimates of the parameters $\hat{\theta} = vec(\hat{\mu}, \hat{\Sigma})$, where the vec() operator stacks the unique elements. The variance matrix of $\hat{\theta}$ is used to draw a simulated θ from a normal distribution with a mean of $\hat{\theta}$ and a variance of V($\hat{\theta}$) (King, *et al*, 2001: 55). The estimate of $\tilde{\beta}$ is computed for the simulated data matrix θ and $\tilde{\varepsilon}$ from the normal distribution. These are substituted into the equation

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i$$

(King, *et al*. 2001: 54)

where

~ indicates a random draw from the appropriate posterior

D_{ij} is a linear function of the other variables

$D_{i,-j}$ are the variables that are not missing data

$\tilde{\varepsilon}_i$ is uncertainty

This procedure is only repeated *m* times, in order to create *m* imputed data sets. As it usually requires only 3-10 imputations, this additional process is generally not computationally expensive. The advantages of EMs are its speed, independence, deterministic convergence rate, and that it can be used with large samples. It functions poorly with small samples (<30), however. Highly skewed categorical data or many variables in comparison with the number of observations may cause bias in the missing data output (King, *et al*, 2001: 55).

EMis (EM with importance re-sampling) improves upon EMs by using the iterative simulation process of importance re-sampling. EMis follows the same procedures as EMs until draws from θ are made. The draws from θ are treated as estimations of the true posterior distribution of the finite sample. The parameters are modified to fit on unbounded scales by using the natural log for the standard deviations and Fisher's *z* for the correlations to make the normal approximation more accurate with smaller samples, such as the one used in this study's database. The program uses and acceptance-rejection algorithm to keeping draws of $\tilde{\theta}$ with probability that are proportional to the Importance Ratio (IR). This ratio is a comparison between the actual posterior and the normal approximation. It can be expressed as:

$$IR = \frac{L(\tilde{\theta} \mid D_{obs)}}{N(\tilde{\theta} \mid \tilde{\theta}, V(\tilde{\theta}))}$$

where IR is the proportional to the posterior distribution of $\tilde{\theta}$ (King, *et al*, 2001: 56). Even in cases of non-normal data, the approximation provides statistically valid results as well as operates quickly. EMis produces multiple imputations from the exact finite sample posterior, as does IP, however EMis is far less expensive computationally (King, *et al*, 2001: 55-56).

EMis was used in the original Amelia I program, but it had faults when dealing with TSCS data sets similar to the one used in this study (Honaker and King, 2007: 3). Often series data forms smooth arcs, but the EMis algorithm would select values far off the line, creating points which made no intuitive sense. Because of this, in the past many researchers in political science have shied away from using MI to fill in missing values, relying upon listwise deletion instead (Horton and Kleinman, 2007: 2). This has made investigating countries like those in the Horn of Africa difficult. Honaker and King developed the Expectation Maximization Bootstrapping (EMB) method for use in Amelia II specifically with political science research using time series data sets in mind (Honaker and King, 2007: 3).

The EMB algorithm uses the EM algorithm to generate an initial estimate of the missing data. Instead of drawing μ and Σ from their posterior density EMB estimates them with a simpler bootstrapping algorithm. The EMB algorithm draws *m* samples of size *n* from the dataset D. For each sample the EM algorithm is run and estimates of μ and Σ are generated. Each set of estimates of μ and Σ are then used to impute the missing observations into their starting positions. The output is *m* multiply imputed data sets. The EMB has the advantage of being faster than EMis, and more faithfully representing the underlying distributions of the data (Honaker and King, 2007:11). See Figure 3-7 for a comparison of posterior estimation by various methods.

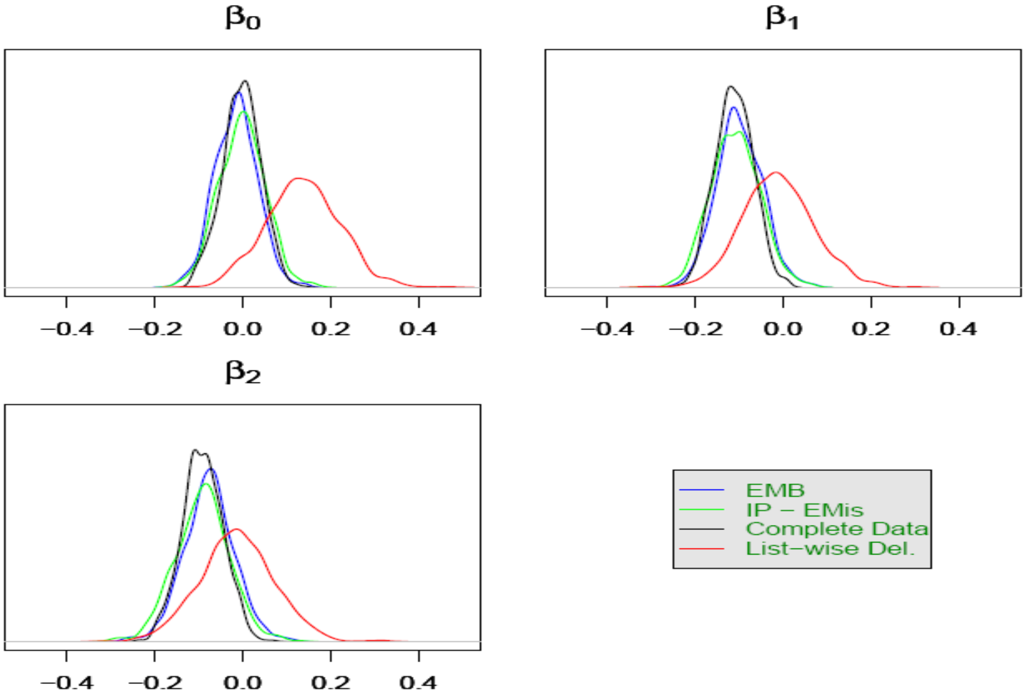


Figure 3-2. Posterior Estimates Using MI Methods (Honaker and King, 2007:11)

The computational speed advantages of EMB are moot until TSCS issues are resolved, however. Amelia II addresses this by recognizing the tendency of some variables to move smoothly over time, to jump between identifying entities (like countries), and for time patterns to differ between those entities. In cases where the data series over time is not smooth the program needs to recognize that a smooth curve is not the best fit for missing data. The algorithm must also allow deviation from a well fitted curve when other highly correlated variables in the model suggest it. Amelia deals with this by allowing the user to designate “time” and “entities”, year and country in the case of this study’s dataset. The program builds a model for each country for each country and each set of variables using locally weighted polynomial (LOESS) regression which builds a model of the time series data that takes autocorrelation into account (Honaker and King, 2007:12). Cleveland, and Cleveland and Devlin provide descriptions of LOESS regression. (Cleveland, 1979) (Cleveland and Devlin 1988). This provides estimates for each missing data point, as well as confidence intervals for each data point. These confidence intervals, if smaller than similar intervals for the originally estimated posterior distribution, now define a new posterior for each data point (Honaker and King, 2007: 13-16). Figure 3-3 shows the difference between *q* order polynomial smoothing and EM confidence intervals. Figure 3-4 shows the difference between polynomial smoothing and LOESS regression. These estimates and confidence intervals represent GDP data for several African countries over time, with the red circles indicating the true data points. It can be seen that use of LOESS regression gives confidence intervals less than 25% of the size of the original EM confidence intervals (Honaker and King, 2007:14)

.

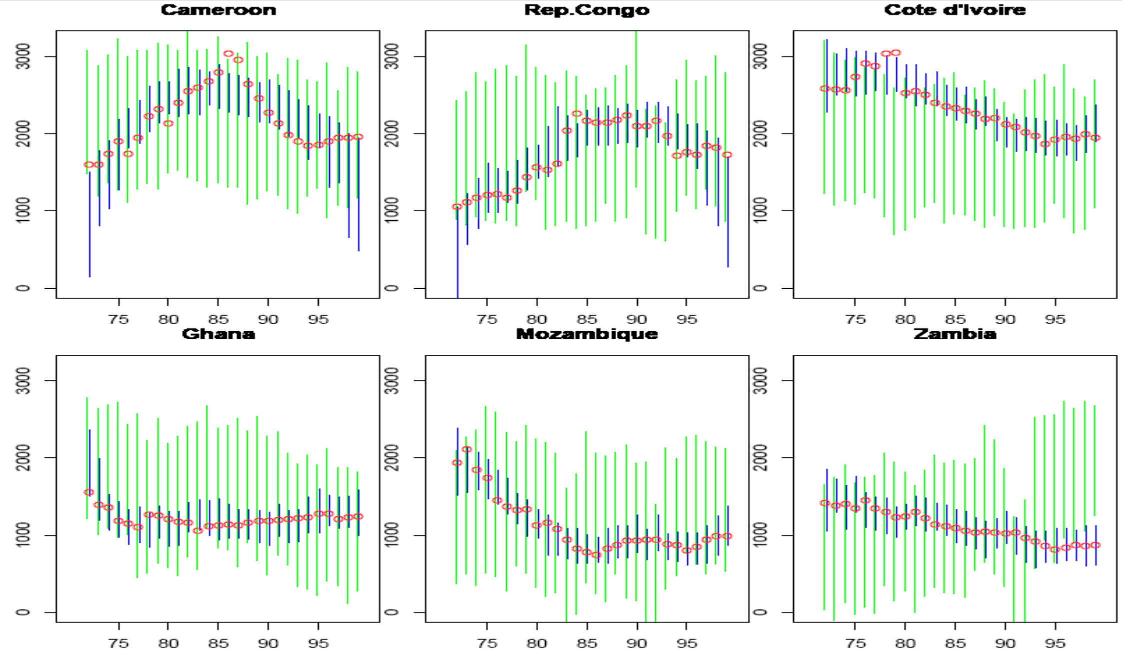


Figure 3-3. Polynomial Smoothing vs. EM estimates (Honaker and King, 2007: 15-16).

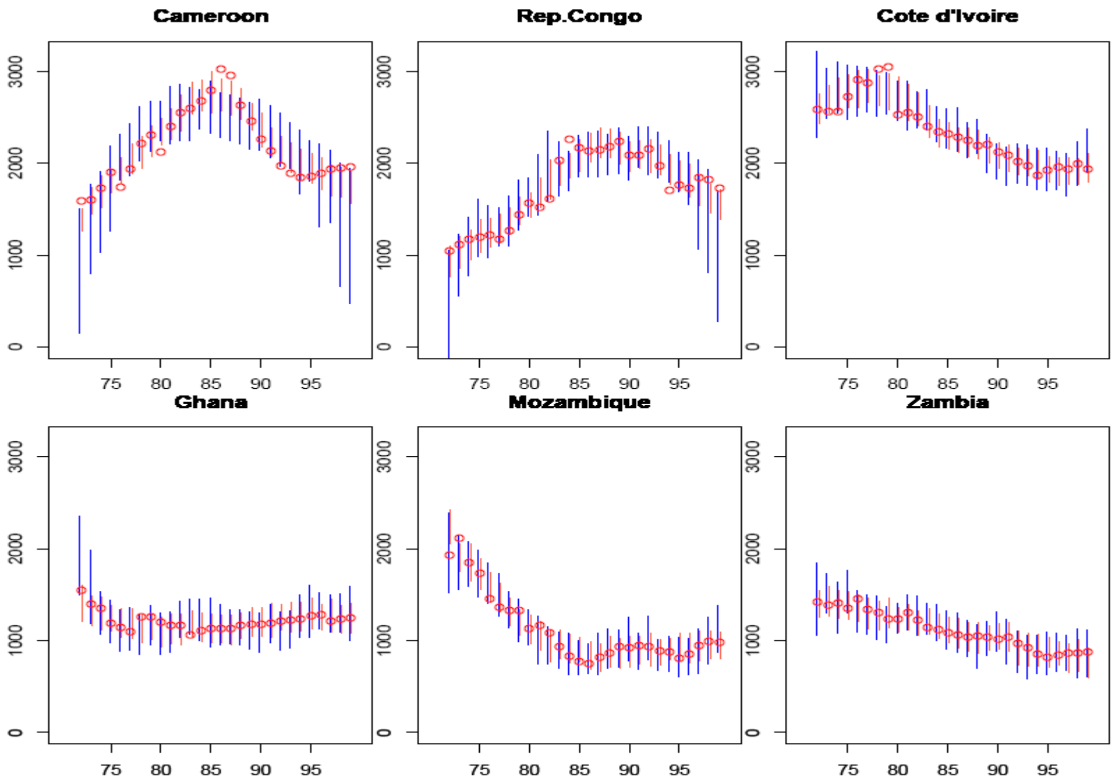


Figure 3-4. Polynomial Smoothing vs. LOESS Regression (Honaker and King, 2007: 15-16).

The user is allowed to define the number of periods examined by the LOESS regression, essentially determining the number of autoregressive terms (Honaker and King, 2007:14). For this study, the maximum of 3 autoregressive terms was selected. This assumption seems justified by the number of the variables examined in Chapter 3.2.6. which show many 3rd order autoregressive terms to be significant. Five multiply imputed data sets were generated. Based on Rubin and Little's formula, the variance of this studies' data sets has only 1.014 times more variance than a similarly generated infinite number of generated data sets (Rubin, 1987: 68).

3.2.8. Imputing Calculated Variables

After creating five MI data sets, some variables representing interaction terms still needed to be calculated for use in the data set using imputed data. In particular, the following were calculated and imputed after multiple imputation: Kilometers of Road per Capita, Water Per Capita and Agriculture Interaction, Land Stress, and the Water / Agriculture / Land interaction term. Definitions of each of these derived variables is provided in Appendix B.

3.2.9. Instability Index

During the earlier phases of this study, some effort was given to developing a single index of instability comprised of battle deaths per capita, refugees per capita, genocide deaths per capita, and undernourishment. The intention was to use it as a continuous dependent variable to use regression models on. The focus of this study shifted away from an instability index when it was decided predictions of individual indicators were feasible and preferable. In addition, issues of validation and interpretability led the research to pursue other avenues. Pilot studies of the index showed a low correlation with their respective Fund for Peace scores. A detailed description of the work done to develop this index is described in Appendices W and X. The results of this index were promising, in that the index did show incidences of the worst humanitarian disasters in the Horn of Africa region over the past 32 years. Figure 3-5 shows the scores for each country.

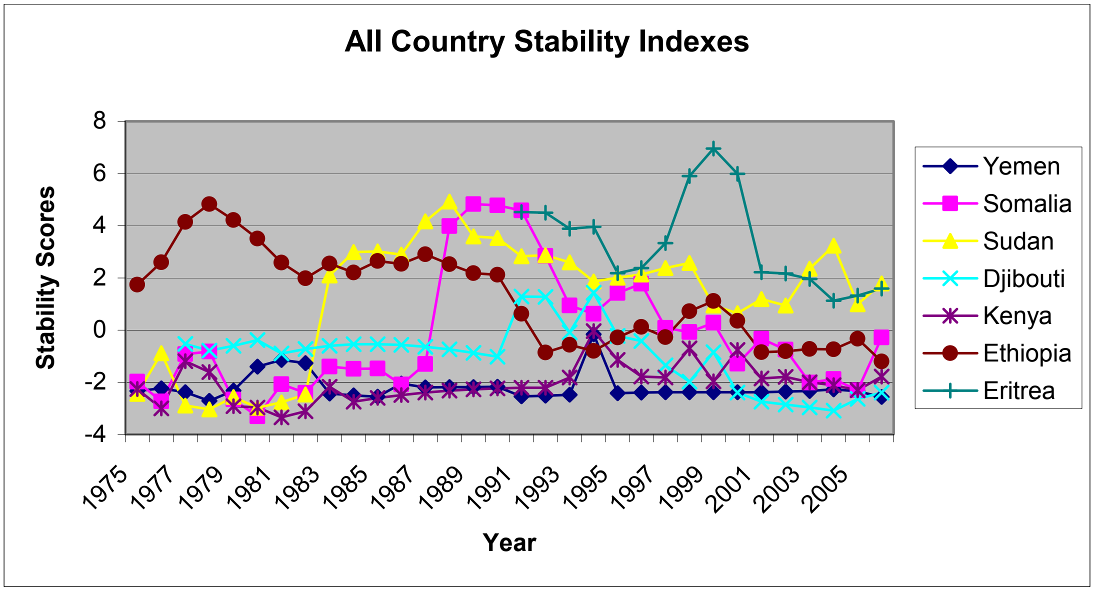


Figure 3-5. Instability Index Scores

The higher the score, the worse the conditions. The peak in 1977 represents Ethiopia during the Ogaden War with Somalia, as well as the terror campaign of Communist genocide and the resulting refugee crisis it generated (Tareke, 2000: 638). The next sharp upturn seen (the yellow line with triangle markers) represents Sudan in 1983 as their civil resumed following an 11 year hiatus over the imposition of Sharia law over the Christian and Animist southern region. This civil war was accompanied by adverse regime change a few years later, and accompanied by genocide by militias and a refugee crisis that remains to this day (de Waal, 2005:24). The magenta line representing Somalia shows a sharp spike between 1987 and 1988, and represents the beginning of the Somali Civil War. During this civil war, genocide, famine, and a refugee crisis were all endemic (UNDPKO, 2008). The last spike represents the Eritrean war with Ethiopia from 1998 through 2000. The severity of this spike is due to Eritrea's relatively small size, the intensity of the conflict, and how the war displaced a larger proportion of its population than other conflicts (GlobalSecurity.org: 2005). Additionally, Eritrea has had the highest malnutrition rates of any of the seven countries in this study over the time period examined. Thus, Eritrea's high instability score, which is passed on per capita numbers, was the most severe according to the instability index. It should be noted that the peaks only occur when a confluence of multiple elements of instability events happen.

The sudden changes in mean of the individual country scores make ARIMA models a poor choice, since they would not predict these changes. Additionally, the fit of models using these scores to hold out data in pilot studies was worse than for individual models (R^2 of .22, compared with between 0.5 and 0.9 for individual instability indicators) The predictability of these scores is examined in Chapter 4; however, given a lack of validation via SME's or sponsoring agencies, the limited predictive abilities of models using the index in pilot studies, and the success of models which provide forecasts of specific instability indicators, the use of this instability index was not further pursued.

3.3 Principal Component Analysis

There are 1176 correlations that describe the relationships between the variables. Given there are only 178 observations when conducting a 4 year forecast, this makes traditional DOE and OLS regression unsuitable, especially considering the high correlation between some variables violated basic assumptions in OLS. Two techniques for generating linearly independent scores representing the structure of the raw data were used in this project: Principal Component Analysis (PCA) and Canonical Correlation. The basic purpose of PCA is to transform an original set of variables into a smaller set of linear combinations which account for most of the variance in the original set (Dillon, 1984: 24). This smaller set can later be used in lieu of the larger set of raw data. This not only generates a smaller set of linearly independent variables, the variables (PCA and canonical correlation scores) are more normally distributed. See Appendix X for a discussion of normal distributions and their tests. It is also possible to interpret each of the these components to understand what effect each variable has on each component or canonical variate, to what degree they effect and which variables group together on a particular component.

Factor Analysis (FA) was considered for use on this study's data set. However, despite some advantages it offered in interpreting loadings, it was rejected because the preferred method of implementation of FA, Maximum Likelihood Estimation, requires the data to be normal, or transformed into normality (Johnson and Wichern, 2002: 426). FA also does poorly dealing with categorical variables. Given the binary data in the data set and the non-removable non-normality of many of the variables, FA was rejected due to the violations of the method's assumptions. PCA, which does not make assumptions regarding the distribution of the data, and was therefore deemed superior for this study (Johnson and Wichern, 2002: 427).

3.3.1. Principal Component Analysis Basic Mechanics

Given the variables $X_1, X_2, ..., X_p$ the objective of PCA is to rotate the original system to using $X_1, X_2, ..., X_p$ as the original coordinate axes (Johnson and Wichern, 2002: 427). These represent the directions of maximum variability and allow the use of a smaller data set. The principal components, PC_p where p is the number of variables, represent the linear combination of the observed variables that accounts for the most possible variance where PC is defined as:

$$PC_{(m)} = w_{(m)}X_1 + w_{(m)2}X_2 + ... + w_{(m)p}X_p$$

where w is a weight chosen to maximize the ratio of the variance of PC_m subject to the constraint $\sum_{j=1}^{p} w_{(1)j}^2 = 1$ (Dillon, 1984: 24-25). In order to reduce the data set, the eigenvalues of the covariance matrix of X, denoted $\sum_{XX}$, are arrange such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ where the covariance matrix is defined by:

$$S = \sum_{XX} = \left(\frac{1}{n-1}\right)\left(X'X - \frac{1}{n}(X'1)(1'X)\right)$$

(Dillon, 1984: 14-15)

Suppose there exists a linear combination

$$\begin{aligned} PC_1 &= w_1^{'}X = w_{1(1)}X_{(1)} + w_{1(2)}X_{(2)} + \cdots + w_{1(p)}X_{(p)} \\ PC_2 &= w_2^{'}X = w_{2(1)}X_{(1)} + w_{2(2)}X_{(2)} + \cdots + w_{2(p)}X_{(p)} \\ &\vdots \\ PC_p &= w_p^{'}X = w_{p(1)}X_{(1)} + w_{p(2)}X_{(2)} + \cdots + w_{p(p)}X_{(p)} \end{aligned}$$

where

$Var(PC_i) = w_i^{'} \sum w_i \quad i = 1, 2, ..., p$

$Cov(PC_i, PC_k) = a_i^{'} \sum a_i \quad i, k = 1, 2, ..., p$ (Johnson and Wichern, 2002: 428)

Thus, the principal components are uncorrelated linear combinations whose variances are as large as possible. The first principal component maximizes $Var(w_1^{'}X)$ subject to $w_1^{'}w = 1$. The second principal component maximizes $Var(w_1^{'}X)$ subject to $w_1^{'}w = 1$ and $Cov(w_1^{'}X, w_1^{'}X) = 0$. For the i^{th} principal component the objective function and the constraints remain the same, with the $Cov(w_i^{'}X, w_i^{'}X) = 0$ for all $k < i$. It is this last property which creates orthogonal scores, and allows those scores to be regressed later without violating the independent variables assumption (Johnson and Wichern, 2002: 427).

This creates a system of p simultaneous linear equations $(S-\ell_1 I)w_1 = 0$ where ℓ is the Lagrange multiplier and chosen such that $|S-\ell_1 I|=0$. In the case of the first principal component, ℓ_1 is the largest eigenvalue of $\sum_{XX}$, and w_1 the associated eigenvector (Dillon, 1984: 30). The contribution of the i^{th} variable to the j^{th} principal component is a number between -1 and 1, and is used to create the p x p loadings matrix. The sign on the individual loading indicates direction. Appendix F shows a sample loadings matrix from multiply imputed data set 5.

It is also possible to perform PCA using the correlation matrix of the raw data set X. This is useful when units within X are on different scales, such as population and a decimal ratio such as gender ratio in primary school. This disparity can influence the derived loadings and scores (Dillon, 1984: 36). Thus, for this study PCA loadings and scores were calculated using the correlation matrix of X, calculated by:

$$R = D^* S D^*$$

where

R is the *p* x *p* correlation matrix of *X*.

D* is a p x p diagonal matrix with elements $\frac{1}{\sqrt{S_{jj}}}$ with $j = 1, 2, ..., p$ (Dillon, 1984: 36)

Finding the loadings matrix of *X* is the same using *R* as it is for *S*, except *R* replaces *S* in all the previous equations. Of note is that the sum of the eigenvalues is *p*, thus the variance explained by each loading is $\ell_{(j)} / p$ (Dillon, 1984: 36). Because this method exists in dimensionless space, it reduces how much variance is loaded at the top end, reducing the technique's effectiveness at reducing the number of variables in a model (Dillon, 1984: 36). Appendix G shows Principal Component Loadings for Data Set MI5.

Component 1 should represent the most variance captured in the model. Note that the largest magnitude loading in any of the columns is -.373 (Foreign Aid as a % of GNI on Component 7). This illustrates one of the drawbacks of PCA, which is the difficulty of interpreting how variables are loaded onto components, and what each component represents. Varimax rotations can be used to help distinguish which variables are most heavily loaded on which components. The mechanics of a varimax rotation will be discussed in section 3.3.2. The eigenvalues associated with the first four components are 9.766, 7.92, 6.303, and 3.94, which captures 20.34, 16.5, 13.13, and 8.22 percent of the total variance. Altogether, the first seven components represent 73.79 percent of the variance in the data set, which shows how a small number of components can capture most of the variance in the complete data set.

The next step in component analysis is to determine how many factors to retain for further analysis. One of the simplest approaches, proposed by Kaiser in 1958, is to retain only those components with eigenvalues greater than 1 (Kaiser, 1958: 190). Another method is called the "scree test" was proposed by Cattel in 1966. This graphical test looks for a knee in the graph of eigenvalues (Cattel, 1966: 145). Figure 3-6 shows the eigenvalues for the data in Data set MI5.

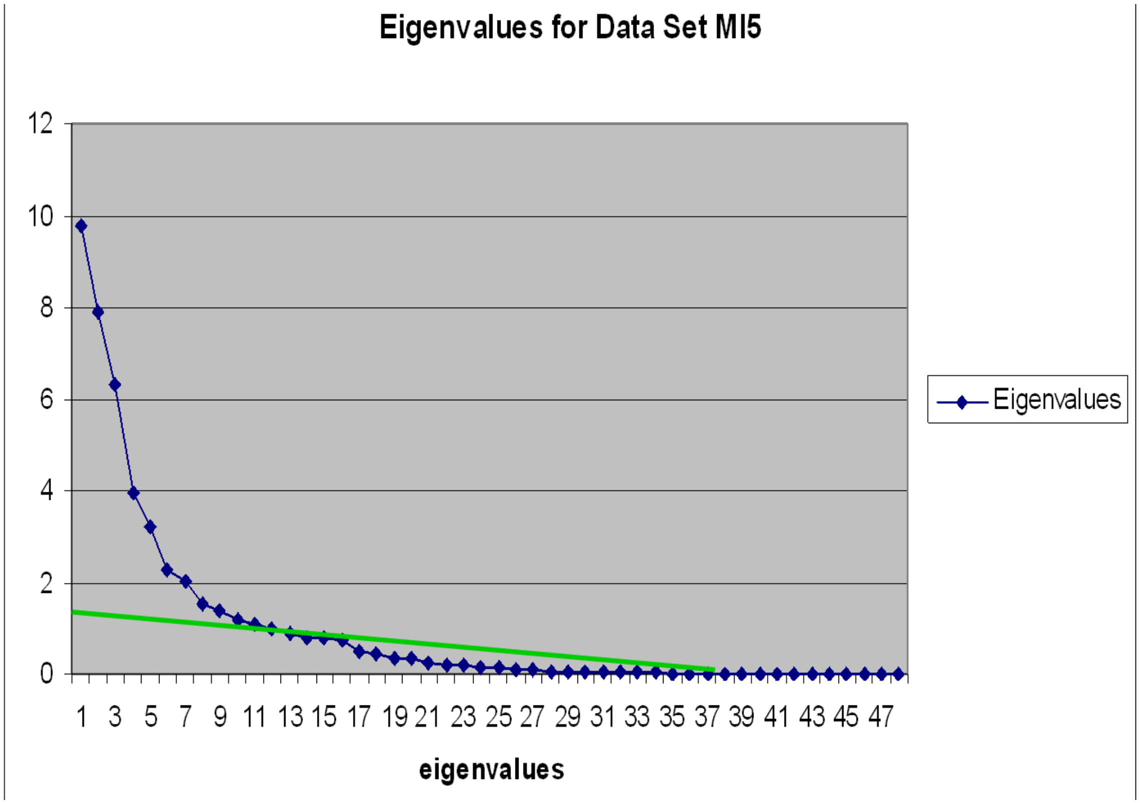


Figure 3-6. Eigenvalues of Data Set MI5

There are two potential points that are evident: data point 11 where the curvature becomes significantly flatter, and at point 17 where there is a sharp drop off noticeable. The green line shows a scree line at point 11. It is not coincidental that the scree line test often suggests an eigenvalue cut off of approximately 1.

The ultimate purpose of PCA in this study was the development of variables based on raw data which were more tractable and appropriate to the forecasting techniques used. For component loadings matrices extracted from variance-covariance matrices the scores are denoted as:

$$
\begin{aligned}
y_{i(1)} &= w'_{(1)}(x_i - \bar{x}) \\
y_{i(2)} &= w'_{(2)}(x_i - \bar{x}) \\
&\vdots \\
y_{i(r)} &= w'_{(r)}(x_i - \bar{x})
\end{aligned}
$$

(Dillon, 1984: 51)

where

x_i is the i^{th} observation vector

r is the number of observations

$\bar{x}$ is the sample mean vector

In matrix form, the component scores can be written as:

$$Y = \left(I - \frac{1}{n} E \right) XA$$

(Dillon, 1984: 51-52)

where

X is the n x p data matrix

E is an n x n matrix of ones

A is the p x r whose columns are the first r eigenvectors of Σ_{XX}

If the scores were calculated using the correlation matrix, the X matrix would be replaced with the standardized score matrix (Dillon, 1984: 51-52). It is worth noting that since the eigenvectors are orthogonal to each other, this causes the component scores to be orthogonal to each other as well, allowing them to be independent. A sample of PCA scores from data set MI5 is shown in Table 3-3. Notice that the magnitude of the scores decreases as the amount of variance explained by the component decreases. In a unitless system, this illustrates how regressing on more component scores has less and less impact on the model. The a proposed interpretation of the components shown in the columns of Table 3-3 is listed in Table 4-2.

Country	Year	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Somalia	1975	-2.3195	-4.4514	0.2248	-2.0976	0.7779	-1.3147	-0.6313
Somalia	1976	-2.118	-4.7234	0.3693	-2.3476	0.8424	-1.3753	-0.7719
Somalia	1977	-2.2896	-4.9117	-0.0609	-2.701	0.8189	-1.009	-0.8731
Somalia	1978	-2.0816	-4.6107	0.2226	-2.4558	1.0414	-1.2435	-0.6335
Somalia	1979	-2.3682	-4.3974	-0.24	-2.7875	1.0001	-1.1449	-0.6589
Somalia	1980	-2.1438	-4.4147	-0.4491	-2.381	1.1071	-0.8421	-0.079
Somalia	1981	-2.1225	-4.3393	-0.5303	-2.6338	0.971	-1.0507	-0.6111
Somalia	1982	-2.1932	-4.3437	-0.5228	-2.4457	1.0692	-0.5035	-0.1739
Somalia	1983	-2.4831	-4.1109	-0.6079	-2.5721	1.3514	-0.9236	-0.1945
Somalia	1984	-2.5557	-3.9641	-1.3373	-2.6595	1.2457	-1.0371	-0.5162
Somalia	1985	-2.3419	-3.9494	-1.5726	-2.2742	1.4659	-0.5845	0.3373
Somalia	1986	-2.513	-4.0015	-1.4576	-2.3502	1.44	-0.562	-0.0533
Somalia	1987	-2.4703	-3.6955	-1.5231	-2.2555	1.4087	-0.6576	-0.2267
Somalia	1988	-2.5631	-3.3953	-1.48	-2.4341	1.6871	-0.6635	-0.5239
Somalia	1989	-2.209	-2.8644	-2.0023	-2.3484	1.7941	-0.875	-0.0577
Somalia	1990	-2.3351	-2.8059	-2.1546	0.135	2.5631	-1.1025	0.3358
Somalia	1991	-1.9539	-2.8175	-1.4936	0.9263	3.0225	-2.352	-1.0283

Table 3-3. Sample Principal Component Scores

3.3.2. Varimax Rotation of Principal Component Loadings

There are numerous types of rotations available for matrices of orthogonal columns. The one most applicable to this study is known as varimax. A rotation is a transformation that rotates a loadings matrix, and is generalized as $A = \Lambda_c T$, where A is the rotated loadings matrix, Λ_c is the loadings matrix, and T is the transforming rotations matrix (Lattin, *et al*. 2003: 145) . Typically, rotations have three common properties:

1. The resulting matrix will still have orthogonal columns.

2. The rotation will not change the communality estimates, but the proportion of each variables variance explained by each component will change.

3. The total amount of variance explained by all the rotated components will not change, but the amount of variance explained individually by the rotated components will (Dillon, 1984: 91).

Varimax is more commonly associated with FA, however, varimax rotated PCA loadings are different from similarly rotated FA loadings. The loadings in PCA are calculated using the correlation matrix, whereas FA finds loadings using either a correlation matrix with its diagonal elements replaced with communality estimates (Principal Factor Method) or more commonly using a maximum likelihood estimation algorithm to generate a model of the population parameters (Dillon, 1984: 81). Using different methods to find the loadings will usually generates different solutions (Dillon, 1984: 85).

The purpose of a varimax rotation is to maximize the variation of the squared component loadings. Normalized loadings are obtained by dividing the each variable's loading by the square root of its communality (Dillon, 1984: 91). This scaling gives each factor or component equal weight in the rotation. The result is to make it easier to identify which variables are loaded on each component.

Kaiser's Varimax rotation is one of the most commonly used (Dillon, 1984: 91). It defines communality as:

$$h_i^2 = \sum_{k=1}^{p} a_{ik}^2$$

where

h_i^2 is the communality of the set of components

a_{ik}^2 is the proportion of variation in variable *i* that is attributable to component j (Lattin, *et al*, 2003: 144)

The varimax rotation seeks to maximize a_{ik}^2 for each column. The variance of column k is given as:

$$V_k = \frac{1}{p}\sum_{i=1}^{p}(a_{ik}^2)^2 - \frac{1}{p^2}\left(\sum_{i=1}^{p}a_{ik}^2\right)^2$$

(Lattin, *et al*, 2003: 145)

Maximizing all of these for each column in the loadings matrix is equivalent to:

$$V = \sum_{k=1}^{c}\ \sum_{i=1}^{p}a_{ik}^4 - \frac{1}{p}\sum_{k=1}^{c}\left(\sum_{i=1}^{p}a_{ik}^2\right)^2$$

(Lattin, *et al*, 2003: 145)

Maximum variance is achieved as one variable is pushed towards -1 or 1 for some component *k*, while all others are pushed (or pulled) towards zero (Lattin, *et al*, 2003: 145). When this is done using normalized loadings, $\frac{a_{ik}^2}{h_i^2}$ is used in place of the simple loadings a_{ik}^2. The illustration Figure 3-7 gives a visual representation of what a varimax rotation does. Notice on the rotated diagram on the right that the distance between the variables (unemployment, AIDS, GDP, and so forth) and the factors (analogous to components) on the x and y axis is reduced to a minimum. These x and y axis can be thought of as two factors or components. In addition, the angle between each variable is unchanged, despite the rotation. This allows the researcher to have a clearer understanding of which variables are really associated with which component (factor). (Lattin *et al*, 2003: 139-140)

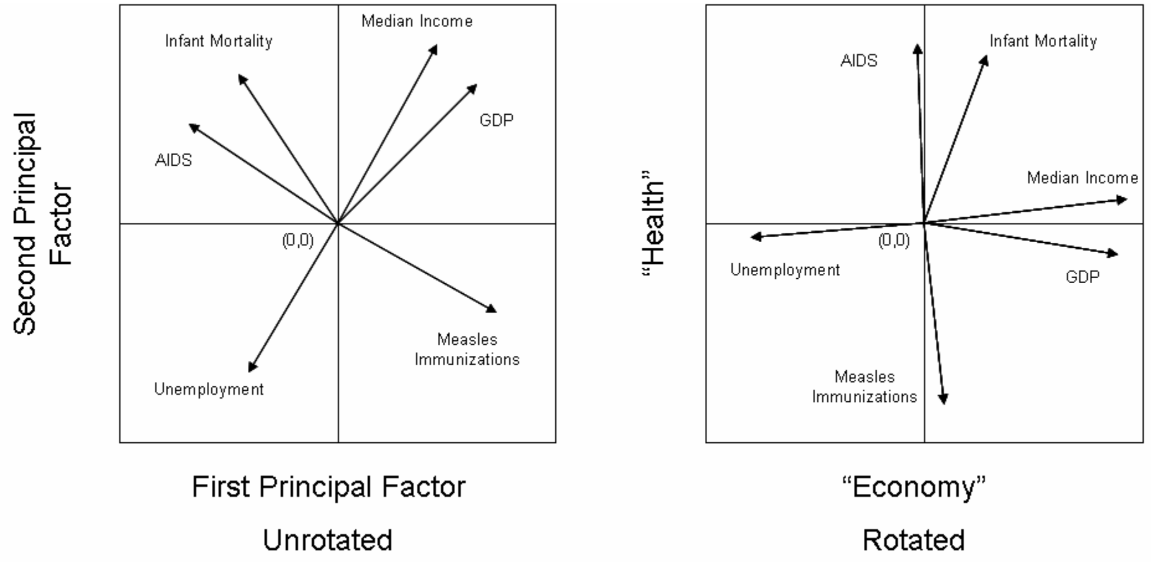


Figure 3-7. Conceptual Diagram of Varimax Rotation (Lattin *et al*, 2003: 140)

Consider the previous example from data set MI5. For illustrative purposes the MATLAB rotatefactor () function was used to rotate the 7 component loadings shown in Table 3-3, and a different set of loadings emerges. Refer to Appendix H: Sample Rotated Loadings Matrix to review the varimax rotated loadings of data set MI5. Referring to Appendix G it can bee seen the best loading has been improved to .441 (Year on Component 4). However, groupings can still be made. For example, the fact that the most highly weighted scores on Component 1, (GDP per capita, Urbanization, Water and Agriculture as a Percent of GDP Interaction, and Relative GDP per capita), suggest that Component 1 might represent how commercialized a country is, and how much their economy has moved away from agriculture, with people moving to the cities where higher wages are available. Indeed, de Waal described this phenomenon in the Sudan as a destabilizing force within the communities that ultimately produced the young men that became the Janjaweed (de Waal, 2006: 60). However, this is speculative, and demonstrates the interpretive nature of using PCA loadings for insight even when varimax is applied. For an interpretation of the full data set and the principal components retained for the study, and a discussion of those principal components, see section 4.2 and Appendix T for the table of rotated factors.

3.4 Canonical Correlation

Canonical correlation is a method for the analysis of interdependence. Much like the previously discussed PCA methodology, canonical correlation also seeks represent a large set of data as a smaller combination. PCA accomplishes this by using loadings to capture the maximum amount of variance in the smallest number of dependent variables. Where canonical correlation differs is it uses two data sets, and the method seeks to maximize the covariance between two sets of linear combinations of the two original data sets (Johnson and Wichern, 2003: 313-314). This type of model has the advantage over PCA scores that PCA scores capture as much variance as possible, including noise, jitter, sampling error, and any other source of variance in the data. Canonical correlation instead tries to capture correlation, or similarities between data sets. Canonical correlation offers the advantage that it does not seek to maximize variance, which might be nothing more than noise, as PCA does. It still suffers some of the same shortcomings as PCA does, namely that interpreting its loadings matrix is somewhat subjective. This section describes the mechanics of canonical correlation, and how it was applied to the data set to produce stability models of the Horn of Africa region.

3.4.1. Canonical Correlation Mechanics

Canonical correlation is a generalization of multiple regression, or OLS (Lattin, *et al*, 2003: 317). However, instead of minimizing the sum of squared deviations the method tries to find a linear combination of the independent data set X which maximizes its correlation with a similar number of dependent variables, Y. In mathematical terms, following the development of canonical correlation in Lattin, *et al*, suppose we have:

$$U = XA$$
$$V = YB$$

where

X is the n x p independent variable data set

Y is the n x q dependent variable data set

n is the number of observations

p is the number of independent variables

q is the number of dependent variables

A and B are n x q and q x q matrices which maximize the correlation between U and V

U and V are n x r matrices where it is assumed $r < p$ (Lattin, *et al*, 2002: 321).

Each of the columns in U are linearly independent of each other, implying $r(U_i, U_j) = 0$ for all $i \neq j$ where $i = 1, 2, ..., q$ (Lattin, *et al*, 2002: 322). The objective is to maximize the correlation between U and V, and can be expressed as:

$$\max \quad r(U,V) = \frac{\text{cov}(U,V)}{\sqrt{\text{var}(U)\,\text{var}(V)}}$$

The covariance matrix between U and V is given by

$$\text{cov}(U,V) = \frac{[U'V]}{(n-1)} = \frac{[U'Y'XV]}{(n-1)} = U'R_{YX}V$$

If U and V are standardized the denominator of the objective function is eliminated, since the transformation results in var(U) = 1 and var(V) = 1 (Lattin, *et al* 2003: 323). Thus, the objective function becomes choose A and B such that $\max \quad B'R_{YX}A$ given

$U'R_{XX}U = 1$ and $V'R_{YY}V = 1$ where

$$\begin{aligned} R_{XX} &= E\{(X-\mu_x)(X-\mu_X)'\} \\ R_{YY} &= E\{(X-\mu_Y)(X-\mu_Y)'\} \\ R_{YX} &= E\{(X-\mu_Y)(X-\mu_X)'\} \end{aligned}$$

(Dillon, 1984: 340)

In order to solve for *A* and *B* the Lagrangian is formed, the first derivative taken, and the derivative is solved for. The Lagrangian is:

$$L = B'R_{YX}A - \frac{\alpha}{2}(B'R_{YY}B-1) - \frac{\beta}{2}(A'R_{XX}A-1)$$

where

α and β are Lagrange multipliers.

Differentiating with respect to A and B gives:

$$\begin{aligned} \frac{\partial L}{\partial A} &= 0 \Rightarrow R_{XY}B - \beta R_{XX}A = 0 \\ \frac{\partial L}{\partial B} &= 0 \Rightarrow R_{YX}A - \alpha R_{YY}B = 0 \end{aligned}$$

(Lattin, *et al*, 2003: 323)

If the two equations above are pre-multiplied by *B*’ and *A*’ respectively, it implies α = r(U,V) = β, where *B* and *A* are the canonical transformation matrices which maximize the correlation between *U* and *V*. To solve via substitution the lower equation isolates B:

$$B = \left[\frac{1}{r(U,V)}\right] R_{YY}^{-1} R_{YX} A$$

Substituting for B into the top equation gives:

$$R_{XY}\left\{\frac{1}{[r(U,V)]}R_{YY}^{-1}R_{YX}A\right\}=r(U,V)R_{XX}A$$

Pre-multiplying by R_{XX}^{-1} and multiplying through by r(U,V) gives:

$$[R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX}]A=r^2(U,V)A\,.$$

This equation is solved by finding A, which is the matrix of eigenvectors of $R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX}$. B can be solved for via back substitution, or it can be solved from the beginning as well, except that it is equal to $R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$ (Lattin, *et al*, 2003: 324). The correlation matrix R can be decomposed into the following, for a better understanding of its sub components:

$$R=\begin{bmatrix}R_{XX} & R_{XY}\\ R_{YX} & R_{YY}\end{bmatrix}$$

3.4.2. Using and Interpreting Canonical Correlation

The interpretation of canonical correlation matrices is very similar to the interpretation of PCA results, where the A matrix is similar to the loadings matrix in PCA, and the U matrix is analogous to scores. These scores were used and tested in lieu of raw data in the forecasting models of instability indicators in the same manner as PCA scores. Canonical correlation scores are independent and more normally distributed than the raw data used in this study. The loadings matrix is useful for the interpretation of the data. The A matrix is used for transforming additional test data into new scores. This property is used both to test the model using hold out data, as well as forecast stability into the years 2009 and 2010 for each of the countries examined in the study.

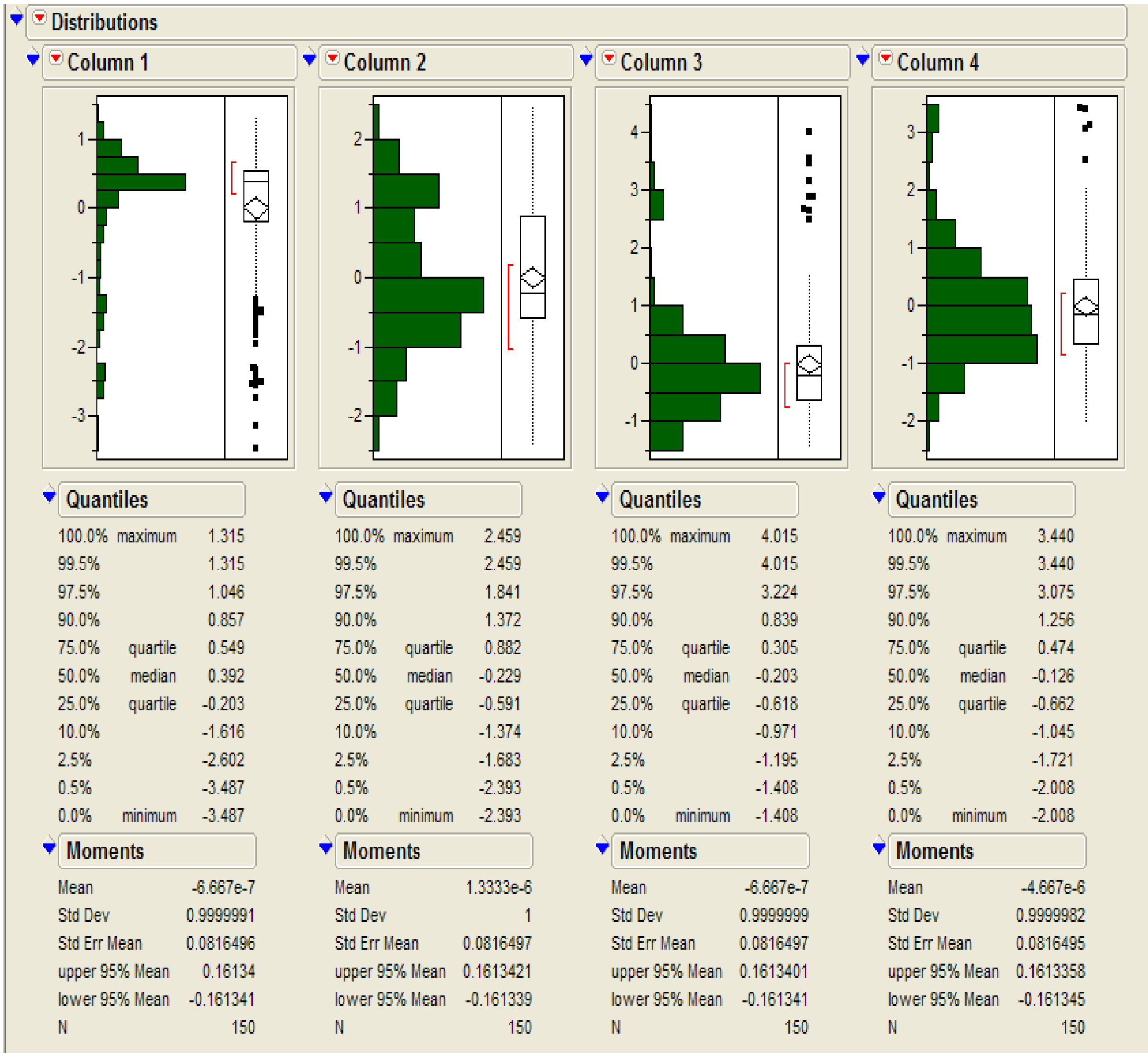
The sample canonical correlations load matrix in Appendix H comes from Multiply Imputed data set 2 (MI 2). The results shown in Appendix H were obtained from MATLAB's built in canonical correlation function, then to find the correlation between X (the raw data set) and U (the canonical correlation scores). One of the outputs is a vector r, which describes the correlation between the j^{th} elements of the U and V matrices. The first column of U and V are always the most strongly correlated, followed by the second as the next most highly correlated pair, and so forth until the last columns are the least correlated (Dillon, 1984: 345). Unmatched columns from U and V, such as the first column in U and the second in V, are uncorrelated. For the sample data in Appendix H, the values of r are 0.9789, 0.915, 0.8389, 0.5966. This also implies that the first column in both the A and U matrices are the most significant to the regression model to be developed later.

Examining the table in Appendix H it can be seen that like PCA loadings, the canonical correlation loadings here are in the range between -1 and 1. It can be seen that undernourishment related data such as previous malnutrition rates, the Ethiopian county identifier, and calories per capita per day were loaded on the first canonical variate. This makes intuitive sense, since later analysis showed that the first variable was the most highly correlated with malnutrition predictions.

The scores in matrix U are linearly independent as a result of being derived from orthogonal eigenvectors. They are also have a mean of zero and a variance of one. Their distribution may not be normal, however. Figure 3-8 shows the distributions of the scores from data set MI2. The results shown in Figure 3-8 come from the JMP statistical package distribution analyzer. Note that the distributions display some skewness, but are all at a minimum mound shaped. They are not bimodal, or multinomial as many of the raw distributions for variable appeared to be. In the final analysis canonical correlation scores were eventually used as independent variables in an OLS regression model, and thus their non-normality was not a violation of assumptions. scores from data set MI2.

Figure 3-8. Sample Distribution of Canonical Correlation Scores Using JMP

The normality, or lack thereof, in the independent variables is not an impediment to OLS regression. What makes canonical correlation scores well suited to regression is their scaling and linear independence. Their scaling makes the interpretation of the regression coefficients somewhat simpler. Their independence satisfies the OLS, DA, and logit regression assumptions of linear independence amongst independent variables.



Quantiles

		Column 1	Column 2	Column 3	Column 4
100.0%	maximum	1.315	2.459	4.015	3.440
99.5%		1.315	2.459	4.015	3.440
97.5%		1.046	1.841	3.224	3.075
90.0%		0.857	1.372	0.839	1.256
75.0%	quartile	0.549	0.882	0.305	0.474
50.0%	median	0.392	-0.229	-0.203	-0.126
25.0%	quartile	-0.203	-0.591	-0.618	-0.662
10.0%		-1.616	-1.374	-0.971	-1.045
2.5%		-2.602	-1.683	-1.195	-1.721
0.5%		-3.487	-2.393	-1.408	-2.008
0.0%	minimum	-3.487	-2.393	-1.408	-2.008

Moments

	Column 1	Column 2	Column 3	Column 4
Mean	-6.667e-7	1.3333e-6	-6.667e-7	-4.667e-6
Std Dev	0.9999991	1	0.9999999	0.9999982
Std Err Mean	0.0816496	0.0816497	0.0816497	0.0816495
upper 95% Mean	0.16134	0.1613421	0.1613401	0.1613358
lower 95% Mean	-0.161341	-0.161339	-0.161341	-0.161345
N	150	150	150	150

3.5. Classification Methods

In previous models various classification methods have been used. In the case of Collier and Hoeffler the purpose of their model was to predict and explain whether a country experienced conflict in a 5 year period based on data from the previous five year interval. In the case of PITF, they were interested in predicting adverse regime changes, ethnic war, civil war, or genocide/ politicide, and classifying a country as unstable if they were experiencing one or more of these events. Both used forms of logistic regression to create classification predictions. In both cases the goal was prediction of which classification countries would fall into at some time hence. To do this, mathematical methods for classification need to be used instead of continuous predictions. This section will discuss Generalized Least Squares (GLS), GLMs, logistic regression, and Discriminant Analysis (DA). GLS is discussed briefly since its use was attempted, along with OLS, to create a better regression models using PCA and canonical correlation scores of instability indicators. Continuous models of battle deaths per capita, refugees per capita, genocides, and undernourishment were attempted using PCA and canonical correlation scores. Only the undernourishment continuous forecasting model had a variance low enough to make it suitable as a forecasting tool for this studies sponsors. Discriminant analysis and logistic regression were used to make discrete forecasting models of the battle deaths, refugee, and genocide data using canonical correlation and PCA scores. These models and their results are discussed in Chapter 4.

3.5.1. Generalized Least Squares (GLS)

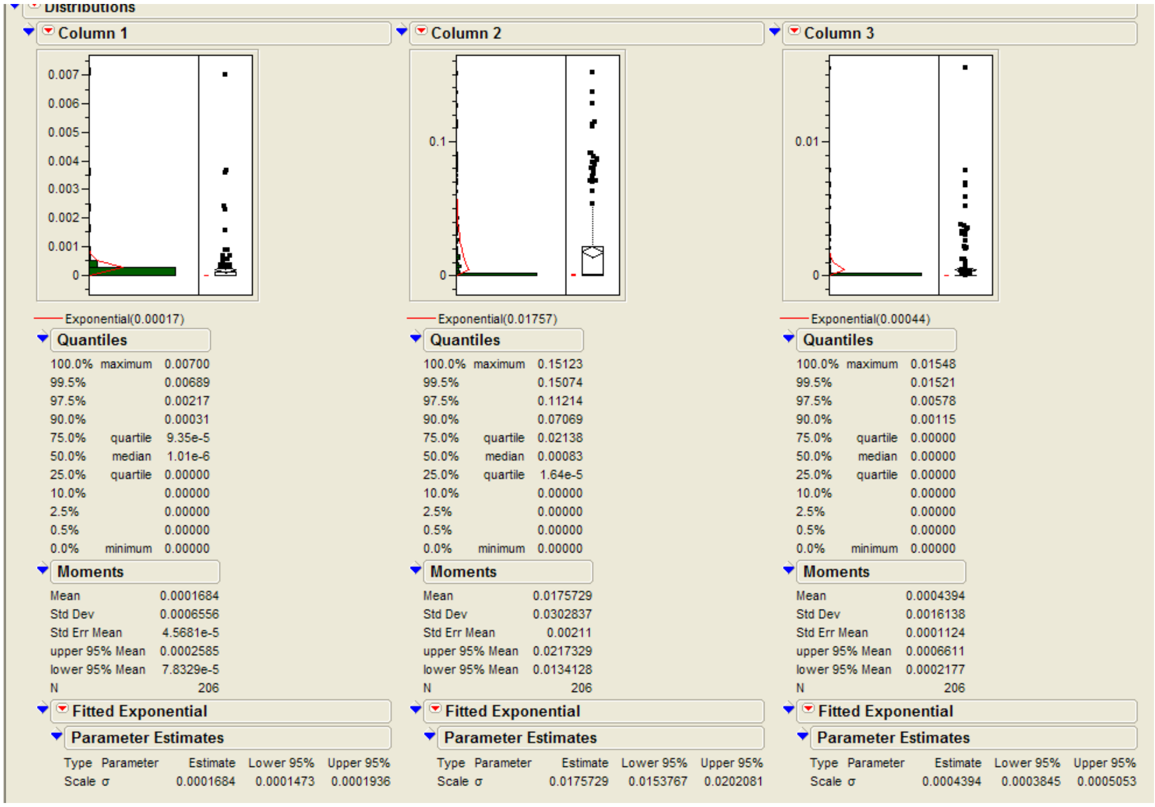
As part of the investigation of the scores produced using the PCA and canonical correlation scores a number of models using GLS were fitted. This section provides a brief description of GLS, since the results produced were poorer than simple OLS regression. Outputs from various GLS model are shown in Chapter 4.

The principal advantage of GLS is that it makes no assumptions regarding the underlying distribution of the dependent variables. This would seem to make it an attractive choice given the non-normality of the battle deaths, refugees, and genocide / politicide deaths, whose distributions could not be transformed into normal ones. Figure 3-9 shows the distributions of the three dependent variables. Note that each of the dependent variables has non-correctable issues with normality. Thus, it suggests use of mathematical models which do not require the dependent data to be normally distributed. OLS assumes the dependent data is normally distributed. GLS is one such family of methods which does not. GLS attempts to compensate for the case where the y_i's are uncorrelated but with unequal variances by including an *n* x *n* matrix V which is defined by $Var(\varepsilon) = \sigma^2 V$ (Montgomery, *et al.*, 2006: 177) . Given V must be non-singular and positive definite, there exists some n x n matrix K such that K'K = V. Thus, given this the following variables are defined:

$$\begin{aligned} z &= K^{-1}y \\ B &= K^{-1}X \\ g &= K^{-1}\varepsilon \end{aligned}$$

such that the normal regression equation $y = X\beta + \varepsilon$ becomes $z = B\beta + g$. This in turn gives the equation $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ (Montgomery, *et al.*, 2006: 179)

The difficulty in using GLS is in finding the structure of the *V* matrix. A variation of the GLS, Weighted Least Squares (WLS), uses clustering. However, the readily available MATLAB DACE (Design and Analysis of Computer Experiments) software package was used to generate GLS models using an array of models for finding the MLE of V. In addition, it allowed for polynomial models. The program uses Kriging surrogates to estimate errors within the model using a correlation matrix R, which is estimated using user defined algorithms (Lophaven, *et al*, 2002: 1). The models generated fittings of the data very precisely; however the prediction results on the hold out data using DACE and GLS had a MSE several times larger than OLS, no matter how the model was specified. The results are shown in Chapter 4. A full description of the methods used in DACE is available in its documentation by Lophaven, *et al*., 2002.



```
Column 1
Exponential(0.00017)
Quantiles
100.0%  maximum   0.00700
99.5%             0.00689
97.5%             0.00217
90.0%             0.00031
75.0%   quartile  9.35e-5
50.0%   median    1.01e-6
25.0%   quartile  0.00000
10.0%             0.00000
2.5%              0.00000
0.5%              0.00000
0.0%    minimum   0.00000
Moments
Mean              0.0001684
Std Dev           0.0006556
Std Err Mean      4.5681e-5
upper 95% Mean    0.0002585
lower 95% Mean    7.8329e-5
N                       206
Fitted Exponential
Parameter Estimates
Type  Parameter   Estimate   Lower 95%  Upper 95%
Scale σ           0.0001684  0.0001473  0.0001936
```

```
Column 2
Exponential(0.01757)
Quantiles
100.0%  maximum   0.15123
99.5%             0.15074
97.5%             0.11214
90.0%             0.07069
75.0%   quartile  0.02138
50.0%   median    0.00083
25.0%   quartile  1.64e-5
10.0%             0.00000
2.5%              0.00000
0.5%              0.00000
0.0%    minimum   0.00000
Moments
Mean              0.0175729
Std Dev           0.0302837
Std Err Mean        0.00211
upper 95% Mean    0.0217329
lower 95% Mean    0.0134128
N                       206
Fitted Exponential
Parameter Estimates
Type  Parameter   Estimate   Lower 95%  Upper 95%
Scale σ           0.0175729  0.0153767  0.0202081
```

```
Column 3
Exponential(0.00044)
Quantiles
100.0%  maximum   0.01548
99.5%             0.01521
97.5%             0.00578
90.0%             0.00115
75.0%   quartile  0.00000
50.0%   median    0.00000
25.0%   quartile  0.00000
10.0%             0.00000
2.5%              0.00000
0.5%              0.00000
0.0%    minimum   0.00000
Moments
Mean              0.0004394
Std Dev           0.0016138
Std Err Mean      0.0001124
upper 95% Mean    0.0006611
lower 95% Mean    0.0002177
N                       206
Fitted Exponential
Parameter Estimates
Type  Parameter   Estimate   Lower 95%  Upper 95%
Scale σ           0.0004394  0.0003845  0.0005053
```

Figure 3-9. Distributions of Battle Deaths, Refugees, and Genocide Deaths Using JMP

3.5.2. Generalized Linear Models

Logistic regression is a specialized case of GLMs. This section provides a brief overall description of GLMs as a background for the description of logistic regression.

GLM is a unifying model because it brings both normal theory linear models and non-linear ones such as logistic and Poisson under the by describing them with a single mathematical formula (Montgomery, *et al*., 2006, 455). The main assumption of GLM is the distribution of the dependent variable belongs to the exponential distribution family, which includes the normal, binomial, Poisson, inverse normal, exponential, Gaussian, and gamma distributions. (Montgomery, *et al*., 2006, 455). These distributions have a general form of:

$$f(y_i,\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + h(y_i,\phi)\}$$

where

ϕ is a scale parameter

θ_i is the natural location parameter

For members of this family:

$$\mu = E(y) = \frac{db(\theta_i)}{d\theta_i}$$

$$Var(y) = \frac{d^2b(\theta_i)}{d\theta_i^2}a(\phi) = \frac{d\mu}{d\theta_i}a(\phi)$$

$$Var(\mu) = \frac{Var(y)}{a(\phi)} = \frac{d\mu}{d\theta_i}$$

(Montgomery, *et al*., 2006, 455)

where $Var(\mu)$ shows the dependence of the variance on the response of the mean. As a result of the last equation:

$$\frac{d\theta_i}{d\mu} = \frac{1}{Var(\mu)}$$

(Montgomery, *et al*., 2006, 455)

The underlying idea behind GLM is to develop a linear model for a function of the expected value of the response variable. Let η_i be a linear predictor where:

$$\eta_i = g[E(y_i)] = g(\mu_i) = x'_i \beta$$

which results in

$$E(y_i) = g^{-1}(x'_i \beta)$$

This last function is the link function, which is simply how the inverse of the transforming function which defines the transformation which translates the dependent data back into it's original distribution. There are many different link functions, each of which is usually associated with a particular distribution. For example, $\eta_i = \mu_i$ for the normal distribution, and η_i is defined as $1/\lambda_i$ for the exponential and gamma distributions (Montgomery, *et al*., 2006, 455). Thus, there are two main parts to a GLM: the distribution of the response and the link function. The link should be treated much like a transformation of the dependent variable except that the link function takes advantage of the natural distribution of the dependent variable (Montgomery, *et al*., 2006, 455).

If $\hat{\beta}$ is the final value of the regression coefficients:

$$E(\hat{\beta}) = \beta \text{ and } Var(\hat{\beta}) = a(\phi)(X'VX)^{-1}$$

(Montgomery, *et al*., 2006, 457)

where V is a diagonal matrix formed using the variances of the estimated parameters in the linear predictor. When using GLM for prediction at some point x_0,

$\hat{y}_0 = \hat{\mu}_0 = g^{-1}(x'_0 \hat{\beta})$ where g is the link function.

3.5.3. Logistic Regression

Logistic regression is a form of GLM which makes categorical predictions based on training data. It can be used in cases where there are more than two possible outcomes, this section however concentrates on the binary case. Logistic regression was used by both Collier and Hoeffler and PITF to build predictive models, and thus this study uses logistic regression for comparison with proposed models as a predictor of battle deaths per capita, refugees per capita, and genocide / politicide deaths (Collier and Hoeffler, 2003: 19), (Goldstone*, et al.,* 2005: 6).

Given the standard regression model of form $y_i = x_i^{'}\beta + \varepsilon_i$, where $x_i^{'}$ and β were previously defined in section 2.4.1.5 and a 0-1 binary response variable y_i which is a Bernoulli random variable, y_i's distribution is defined as:

$$P(y_i = 1) = \pi_i$$
$$P(y_i = 0) = 1 - \pi_i$$

(Montgomery, *et al*., 2006, 429)

Since the expected value of ε_i is zero, it implies $E(y_i) = x_i^{'}\beta = \pi_i$; the function simply states that given some observation x_i the probability the response takes on a value of one is π_i, where $0 \le \pi_i \le 1$. The actual values of ε_i can only be 1, -1, or zero since

$$\varepsilon_i = 1 - x_i^{'}\beta \text{ when } y_i = 1$$

$$\varepsilon_i = -x_i^{'}\beta \text{ when } y_i = 0$$

Thus, errors in logistic regression are not normally distributed. In addition, the error variance is not constant since $\sigma_{y_i}^2 = E(y_i)[1 - E(y_i)]$, indicating the variance of the observations is based on the mean (Montgomery, *et al.*, 2006: 429).

When the response in logistic regression are binary it generally results in a response curve that is non-linear, monotonically increasing or decreasing, and S-shaped or reverse S-shaped. Figures 3-10 and 3-11 show examples of these curves.

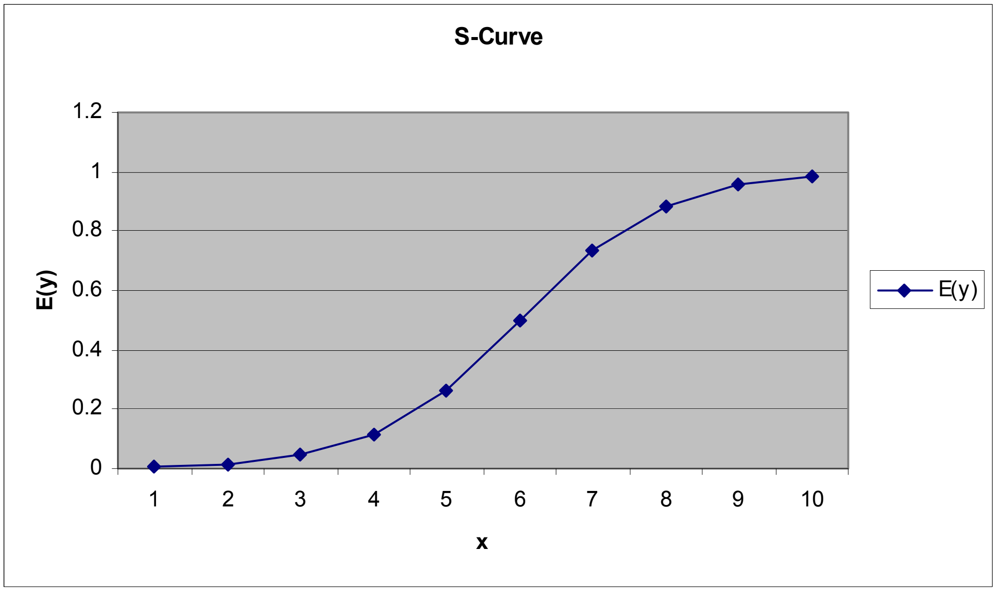


Figure 3-10. Monotonically Increasing S-Curve with $0 \leq E(y) \leq 1$ (Montgomery, *et al.*, 2006, 429)

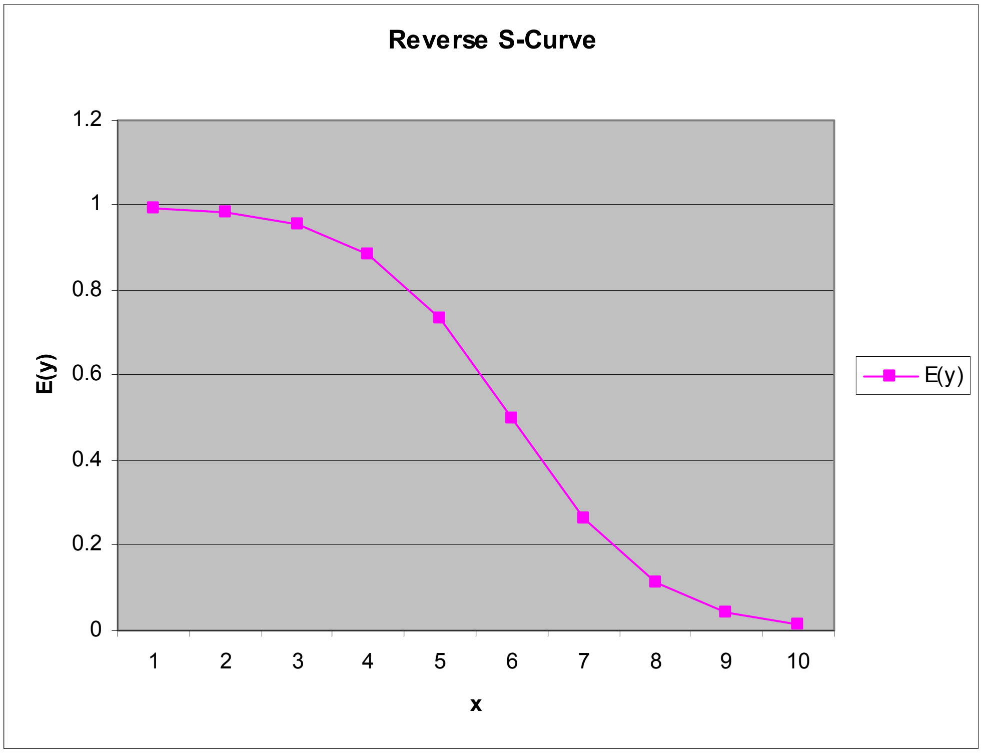


Figure 3-11. Monotonically Decreasing Reverse S-Curve with $0 \le E(y) \le 1$ (Montgomery, *et al*., 2006, 429)

These functions are of the form:

$$E(y) = \frac{1}{1 + \exp(-x'\beta)}$$

Linearizing the logistic response function is accomplished via the link function

$$\eta = \ln \frac{\pi}{1 - \pi}$$

where $\eta = x'\beta$ is linear predictor (Montgomery, *et al*., 2006: 430).

The logistic regression calculations used in this study were done in MATLAB, which uses the maximum likelihood estimator to find the parameter β. Since the observations are Bernoulli, the probability of each is $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, $i = 1, 2, ..., n$ where y_i is 0 or 1. Given independent observations the likelihood function is:

$$L(y,\beta)=\prod_{i=1}^{n} f_i(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

(Montgomery, *et al.*, 2006, 430)

When converted to the log likelihood function this becomes:

$$\ln\ L(y,\beta)=\sum_{i=1}^{n}\left[y_i\ \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right]+\sum_{i=1}^{n}\ln\ \left(1-\pi_i\right)$$

Given $1-\pi_i=\left[1+e^{x_i'\beta}\right]^{-1}$ and $\eta_i=\ln\ \left[\frac{\pi_i}{1-\pi_i}\right]=x_i'\beta$, the log likelihood function can be re-written as :

$$\ln\ L(y,\beta)=\sum_{i=1}^{n} y_i x_i'\beta-\sum_{i=1}^{n}\ln\ \left[1+e^{x_i'\beta}\right]$$

The Maximum Likelihood Estimates (MLEs) of the vector of coefficients β can be solved for using numerical search methods, or Iteratively Re-weighted Least Squares (IRLS). This is how MATLAB calculates the MLE the vector of coefficients. A detailed description of IRLS method can be found in Appendix C.14 of the Montgomery text (Montgomery, et al., 2006: 573).

If $\hat{\beta}$ is the final estimate of β and the assumptions are correct:

$$E(\hat{\beta})=\beta \text{ and } Var(\hat{\beta})=(X'VX)^{-1}$$

Where V is a n x n matrix containing estimates of variance of each observation on each diagonal element, defined as $V_{ii}=n_i\hat{\pi}_i(1-\hat{\pi}_i)$ (Montgomery, *et al.*, 2006: 431).

3.5.4. Discriminant Analysis

The purpose of Discriminant Analysis (DA) is to classify observations into mutually exclusive and exhaustive groups on the basis of a set of independent variables.

When used to classify observations into a binary category, DA acts like logistic regression as a tool for classifying objects into one of two groups, although logistic regression creates a function or functions and scores to separate observations into groups. The way this project defined groups was by establishing number of battle deaths, refugees, or genocide deaths as a threshold, and dividing countries by whether they had more of these four years in the future than the selected threshold, or less. This was done for both the raw instability indicator data, and the per capita data. For example, if the refugee threshold was set at 2000, then Yemen's total refugees for 1979 would be examined. Supposing that Yemen had 15000 refugees in 1979, then Yemen's 1975 country data would be put in the group for data sets representing an outcome of more than 2000 refugees in a year. Conversely, if it had been less than or equal to 2000, then the 1975 data for Yemen would have been placed in the other group representing less than or equal to the threshold.

DA's overall goal is to discriminate between two groups of objects, and minimize the classification error rate. This is accomplished by maximizing the between group variance relative to the intra-group variance (Dillon, 1984: 360). In simplistic terms, DA assigns a score that is a weighted average of the values on the set of independent variables. Given this score, it can be transformed into a probability that it belongs to each of the *a priori* groups. The basic goal of DA can be expressed as:

$$\Delta = \frac{b'\mu_1 - b'\mu_2}{b'\sum b}$$

where

Δ is the maximized apparent distance between the two groups

μ_i is the mean of each of the two groups, such that $\mu_i = x_i^{'} = (\overline{x}_{i1}, \overline{x}_{i2}, ..., \overline{x}_{ip})$

b is the vector of weights which maximizes Δ

Σ is the pooled covariance matrix found by:

$$S \approx \Sigma = \frac{1}{n_1 + n_2 - 2}(x_1^{'}x_1 + x_2^{'}x_2)$$

(Dillon, 1984: 364-365)

where x_i is the *p* x n_i matrix of observations corresponding to each of the two *a priori* groups. Thus, by substitution it can be shown the estimate of *b*, $\hat{b}$, is found via:

$$\hat{b} = S^{-1}(\overline{x}_1 - \overline{x}_2)$$

where S^{-1} is the inverse of the pooled covariance matrix (Dillon, 1984: 364-365).

In this study DA and logistic regression were used on the data for battle deaths, refugees, and genocide deaths and their per capita equivalents. In each case, each method was tested on each multiply imputed data set using a range of values of the dependent variables as a dividing point for the two *a priori* groups in the training data. For example the threshold, or dividing point, for battle deaths was first set at 0. For DA, all countries which had 0 battle deaths four years after the data in an observation were separated into one group, and those with one or more battle deaths put in the second group. For the logistic regression model, countries with 0 battle deaths 4 years hence were given a dependent variable score of 0, and those with 1 or more battle deaths were given a score of 1. A model was created using both DA and logistic regression, and then used on the hold out data set to check its predictive ability. The results were recorded, and then the process started over again using a new dividing point, in this case iterated so that 1 or less resulted in placement in one group or a score of 0, and 2 battle deaths or more resulted in placement in the other group or a score of 1. The MATLAB code used to build and test the DA and logistic regression models iteratively can be found in Appendix J.

DA distances can be tested for significance. This is useful for describing the statistical significance of both the model, as well as examining the significance of each variable within the model. The following discussion describes the procedure for testing whether between group distances are significant. Mahalanobis's generalized distance is defined as:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

This distance is used to find a test statistic, the Hotelling's two-sample T^2 statistic, which is F-distributed and defines as:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

Given this, $H_0: \mu_1 = \mu_2$ can be rejected at the significance level α if:

$$\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 > F_{\alpha;(p, n_1 + n_2 - p - 1)}$$

(Dillon, 1984: 368)

This method of statistical testing can be applied both to the whole model and to each variable one at a time (Dillon, 1984: 366-368).

Another method for examining the importance of individual variables within the DA model is the use of discriminant loadings. These give a simple correlation of a variable with the discriminant function. DA loadings are found by first re-scaling the discriminant weights with:

$$b_j^* = C\hat{b}_j$$

where C is a vector containing the square roots of the diagonal elements of S. After this, the actual load is found using:

$$\hat{\ell}_j = R \cdot b_j^* (\hat{b}_j \cdot \text{cov}(X) \cdot \hat{b}'_j)^{-1/2}$$

where R is the correlation matrix of the complete set of training data and X is the centered data. Because this method captures common variance among the predictors, it is the preferred method of evaluating the influence of individual variables on the model (Dillon, 1984: 373). The loadings, $\hat{\ell}_j$, were used to make inferences regarding the effects variables had on the model.

There are two assumptions with DA: equal variance-covariance structures, and multivariate normality. The former is less of a problem to this study, since the groupings of the training data varies from one iteration to the next, the variance of the data points between sets can be assumed to be approximately equal over the entirety of the points examined. The multivariate normality assumption is not met, however. Much of the data in this study is binary, or has a distribution which cannot be transformed into a normal one. As a consequence, tests of significance and classification errors may be biased. However, the purpose of this study is to examine different predictive models in a broad spectrum of conditions, and a model which displays strong predictive abilities despite violating assumptions is still a useful tool. DA has often been used in the past on non-normal data. The effect of non-normality can be seen in this study manifested as errors that strongly favor false positives, or false negative results. Biased errors may actually be preferable, though. This will be shown and discussed in Chapter 4.

3.6. Evaluating and Comparing the Models

During the development of the competing models, there must be standards by which to evaluate them against each other, as well as previous models discussed in Chapter 2. Several of the models investigated used OLS regression for developing a vector of coefficients, a number of standard measures of fit exist. R^2 and R^2_{adj} have already been described, but other statistics such as root mean square error (MSE), and R^2_{pred}. In cases where discrete models were developed, forecast error, or Apparent Error Rate (APER) on hold out data was the primary method of comparing the models. Additionally, the types of error (false positives or false negatives) was a factor in comparing the discrete models. Another objective is to compare these new models to existing ones, particularly those developed by the CAA, Collier and Hoeffler, and the PITF. Comparing the models developed in this study with previously developed models poses a particular problem, since Collier and Hoeffler as well as PITF chose to model instability in a different way, and offered different lead times.

Following whole model trials, the data was divided into predictor and predicted sub-sets. The predicted sub set was the previous 4 years for each country, i.e. 2003-2006 and 4 year lagged data from each country. This reduces the number of data points for creating a model 178 to 150, or a hold out of approximately 16%. This is smaller than was used by the PITF, however the number of data points available to build their model was approximately 50 times larger (Goldstone, *et al*, 2005: 7). The variables used in the models are described in the Chapter 4. The standard by which the continuous models are measured in terms of prediction utility for this study are RMSE and R^2_{pred}. This last statistic is based on the PRESS statistic, and defined by:

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

$$R^2_{pred} = 1 - \frac{PRESS}{SS_T}$$

(Montgomery, *et al*, 2006: 125))

where

y_i is the actual t-score for data point *i* in the hold out set

$\hat{y}_i$ is the predicted t-score for data point *i* in the hold out set

R^2_{pred} and MSE were used to evaluate both the whole model approach, and the predictive model tests. Since the columns of scores were orthogonal, correlation and multicollinearity were not issues for the PCA and canonical correlation scores, however highly correlated data and multicollinearity were issues with the formation of the scores. Unnecessary data was identified using the canonical correlation loadings, and the data model created without those variables. These same columns were also eliminated when finding PCA scores and loadings.

The two most commonly seen models in the literature have been Collier and Hoeffler's, and the PITFs. The former is less a predictive model, than one which explores and compares social and economic factors as potential causes of instability, namely armed conflict. Thus, the pseudo-R^2 values for their predictions does not exceed 0.30. It should also be noted that their window of prediction, rolling 5 year intervals, measures instability in large blocks, rather than on a year by year basis (Collier and Hoeffler, 2001: 26).

The PITF model is primarily concerned with predictive ability, and uses a stepwise logistic on two year lagged data to give a binary prediction of stable or unstable. Another key difference between PITF and this study is PITF was only concerned with predicting the onset of instability. Data that occurs within five years after the end of an instability event was not part of their model (Goldstone, *et al*: 2005: 7-10). Thus, a one to one comparison between models is not possible, since the outcome spaces differ. However, the PITF Phase V report includes prediction error rates, which are compared the prediction error rates of the discrete models developed by this study using logistic regression and DA. Prediction error is defined as:

$$APER = \frac{\sum Forecast \quad Errors}{Number \quad of \quad Hold \quad Out \quad Po\text{int}\,s}$$

(Goldstone, *et al*: 2005: 10)

A secondary purpose of this study was to investigate which raw variables are connected with future instability incidents in the HoA. For both PCA and canonical correlation this was be done by examining and comparing the varimax rotated loadings matrix of the PCA scores and the canonical correlation loadings matrix, both of which show variable loadings on the their respective scores matrices. Despite the fact that since some of the data was multiply imputed, and each data set unique, there was little difference between the loadings for each data set. As with most loadings matrices, the labeling of the pseudo-variables was subject to interpretation. A detailed description of the model and analysis of the data follows in Chapter 4.

3.7. Describing the Models

The methods and techniques described in Sections 3.4 and 3.5 were used to build models of the four instability indicators. Numerous models, both continuous and discrete, were tested and evaluated. The best results and model parameters are described in Chapter 4, and records of most of the models tested are listed and described in the Appendices.

The data reduction techniques described in Section 3.4, principal component analysis and canonical correlation, were used to generate scores which better met the assumptions of the techniques described in Section 3.5. The data was standardized prior to using PCA and canonical correlation to generate the scores. These PCA and canonical correlation scores were tested in OLS, GLS, logistic regression, and discriminant analysis models. After generating models and evaluating them in terms of fit, variance, utility, apparent error rates, and types of error generated one model for each of the 4 instability indicators was accepted for use in further study of the key variables, comparison with previous models, and to make forecasts for 2007-2010. The following listing gives a brief description of each of the models and the scores used to generate them:

1. Continuous forecasting model of undernourishment for each country in the Horn of Africa region using canonical correlation scores and Ordinary Least Squares regression.

2. Discrete forecasting model of battle deaths per capita for each country in the Horn of Africa region using principal component scores and logistic regression.

3. Discrete forecasting model of refugees per capita for each country in the Horn of Africa region using principal component scores and logistic regression.

4. Discrete forecasting model of genocide and politicide for each country in the Horn of Africa region using principal component scores and Discriminant Analysis (DA).

A detailed description of each of these models, their fit, results, interpretation, and forecasts is given in Chapter 4.

3.8. Chapter Summary

This chapter outlined the data collected and used in the study. It described the methods used to impute and simulate the missing data. PCA and canonical correlation were described as methods for creating independent data to be regressed. Several different methods of regressing continuous dependent data were described, along with logistic regression for creating models of categorical data. DA was described an alternative method of building a model of categorical data. Finally, several methods evaluating continuous and categorical models were described.

4 Analysis Results

4.1. Introduction

This chapter begins with proposed interpretations of the canonical correlation and Principal Component Analysis (PCA) scores used as variables in the forecasting models developed by this study. It then describes and discusses the continuous model of undernourishment and the discrete models of battle deaths per capita, refugees per capita, and genocide. The variables significant to each of these models are examined. The discrete models are compared with other similar models, and forecasts for the years 2007 through 2010 are provided based on the models put forward by this study. Significant contributions to the field and suggestions for further research are discussed in Chapter 5.

4.2. Principal Component and Canonical Correlation Loadings

For each of the of the models developed for this study, either canonical correlation scores or principal component scores calculated via the varimax rotated loadings matrix were used as data points. 54 variables were used to generate both the canonical correlation scores and the principal component scores. These variables are listed in Appendix K. The principal component loadings matrix was calculated using data from 1975 through 2002.

Although the models were developed prior to the analysis and interpretation of the loadings, they are presented here to facilitate the reader's understanding of what each of the principal component and canonical correlation scores represent as the models of instability indicators are presented in turn. There were four canonical correlation scores generated, however only the loadings of the first three are shown, since the fourth did not prove to be significant to the model of undernourishment at a α of .95. Table 4-1 shows the variables with the highest loadings on the first three canonical correlation scores.

Independent Canonical Correlation Score					
1		2		3	
Variable	Loading	Variable	Loading	Variable	Loading
4 Year Lagged Refugees	-0.6659	Trade Openness	-0.5571	4 Year Lagged Genocide Deaths	0.5591
4 Year Lagged Undernourishment	-0.6051	Forrested Land	0.5005	Land Stress	0.5242
Calories	0.6005	Agriculture as % GDP	0.4527	Forrested Land	0.5226
4 Year Lagged Battle Deaths	-0.5622	Military as % GDP	-0.4385	Economic Discrimination	0.5242
		Land Stress	0.4358	Water / Agri / Land Intteraction	0.4475
				Bad Neighbors	0.4048
R-squared	0.9763	R-squared	0.9171	R-squared	0.8387

Dependent Variable Canonical Loadings				
	Canonical Variate			
	1	2	3	4
Undernourishment	-0.8794	-0.4357	0.1849	-0.0515
Battle Deaths	-0.5591	0.2275	-0.0283	0.7968
Refugees	-0.7859	0.5995	-0.1434	-0.05
Genocide / Politicide Deaths	0.0597	0.4004	0.914	-0.0273

Table 4-1. Highest Canonical Correlation Scores Loadings

The full set of independent variable loadings is shown in Appendix N. Undernourishment, battle deaths, refugees, and genocide are listed in both the dependent and independent variables since existing levels of instability were part of the data used to predict future instability, i.e. 1975 figures on undernourishment was part of the data used to predict undernourishment in 1979, and so forth. The loadings represent the correlation between the input raw data and the output canonical correlation scores. From the table of dependent variable loadings it can be seen that undenourishment is most heavily loaded on the first two dependent variable canonical variates. Note that the first set of canonical correlation scores (or canonical variates, since these scores are treated exactly as a variable would be using OLS, logistic regression, or DA) generally reflect existing conditions of current instability and undernourishment. The second centers on economic reliance on agriculture, scarcity of arable land, and the countries' openness to trade. The third canonical score seems to be centered on genocide and the associated variables.

The same 54 variables were used for PCA analysis as were used for canonical correlation, and are listed in Appendix K. The raw data was handled slightly differently for PCA than canonical correlation, however. PCA was used on the entire set of entering data from 1975 through 2002. The justification for this is based on how hold out data would be used if a researcher in 2002 wished to build a predictive model similar to the one used in this study. It is assumed that they would likely use all the data available up to 2002 in order to provide a more complete model of the data's structure. After PCA loadings and scores were found for the combined training and hold out set, the scores were separated into their normal training and hold out sets. The variances were examined and 13 components retained based on a scree line test, and the consideration for which eigenvalues exceeded 1. Figure 4-1. shows the eigenvalue plot. Using 13 components accounted for 88.45% of the total variance in the data. The 54 *x* 13 loadings matrix of retained components was rotated using a varimax rotation and then the rotated loadings matrix was multiplied by the standardized raw data set to generate a set of rotated scores.

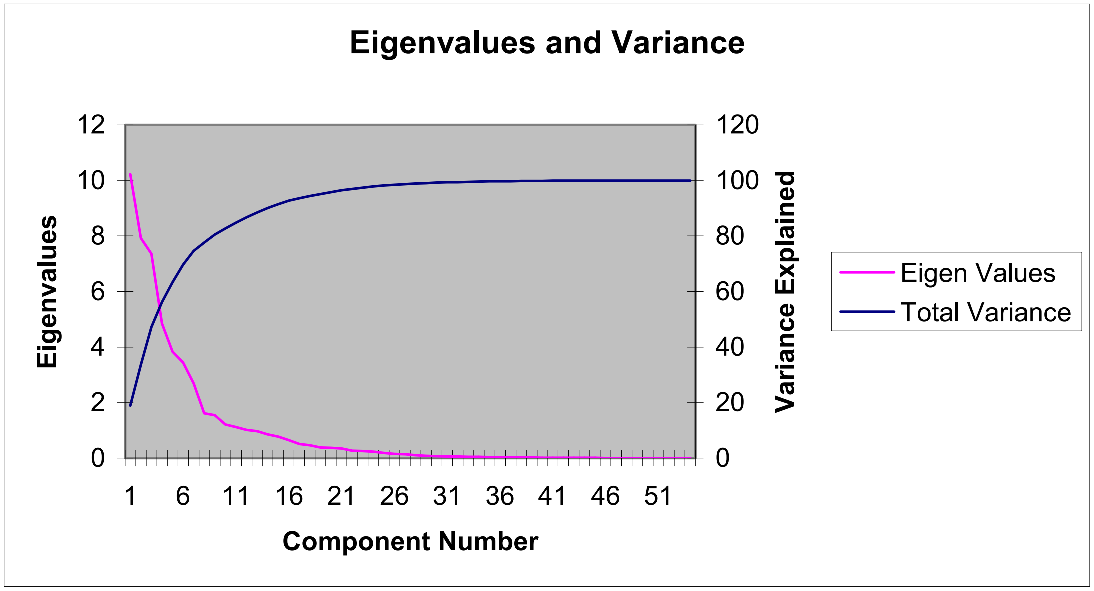


Figure 4-1. Eigenvalues and Variance Explained Using PCA

The rotated loadings of the 13 retained components are shown in Appendix T. Based on these rotated loadings, a subjective interpretation of the 13 principal components is hypothesized and shown in Table 4-2.

PC #	Interpretation	Eigenvalues
PC1	Wealth and Urbanization	11.3773
PC2	Fractionalization	8.6745
PC3	Agricultural Reliance / Stress	7.1561
PC4	Total Population, Ethiopia	4.9866
PC5	Time Effects	3.6205
PC6	Military / Health Spending	3.1939
PC7	Economic Integration into Global Market	2.1737
PC8	Long Term Government Stability	1.7035
PC9	Weak Government / Existing Instability	1.6422
PC10	Change in Infant Mortality Rate	1.3194
PC11	Political Discrimination / Exclusion	1.1372
PC12	Improving Quality of Life	1.0995
PC13	Transition Government	0.9948

Table 4-2. Proposed Interpretation of Principal Components

These interpretations are shown to assist the reader in understanding which types of variables are affecting the models of battle deaths per capita, refugees per capita, and genocide as they are presented in this document. The independent data in these principal components is lagged 4 years behind the dependent data, and includes lagged undernourishment, battle deaths, refugees, and genocide deaths.

4.3. The Forecasting Models

This section presents a forecasting model for each of the indicators used to quantify instability. The models can be summarized as:

1. Continuous forecasting model of undernourishment for each country in the Horn of Africa region using canonical correlation scores and Ordinary Least Squares regression.
2. Discrete forecasting model of battle deaths per capita for each country in the Horn of Africa region using principal component scores and logistic regression.
3. Discrete forecasting model of refugees per capita for each country in the Horn of Africa region using principal component scores and logistic regression.
4. Discrete forecasting model of genocide and politicide for each country in the Horn of Africa region using principal component scores and Discriminant Analysis (DA).

4.3.1. Continuous Model of Undernourishment

Continuous models of each of the instability indicators were explored as part of this study using different scores, variables, and modeling techniques. Of them, a model of undernourishment using canonical correlation scores and OLS regression proved to be of the greatest accuracy, both in terms of variance, measured as Root Mean Square Error (RMSE), and in terms of variance accounted for in terms of R^2, R^2_{adj}, and $R^2_{\Pr ed}$. Results for the continuous models of other instability indicators are shown and discussed in Appendix Z.

Five data sets were created for this study using multiple imputation via the Amelia II software. The first of the five, Multiply Imputed Data Set (MI 1), was used for development of this model, and the MATLAB "canoncorr" function was used to generate the canonical correlation scores. Later tests showed use of the other data sets did not significantly alter the model parameters. The MATLAB function was modified to allow the creation of new scores using hold out data for the purpose of evaluating the method's predictive value. Two variables were removed from the training data since they caused the correlation matrix which generates the loadings be singular: Partial Democracy without Factionalism and Religious Fractionalization. Appendix K shows the variables used to generate the training independent data matrix. Training data for each country from 1975 until 1998 was placed in one group of variables and used as the independent data set. Undernourishment, battle deaths, refugees, and genocide data from 1979 through 2002 was placed in the other and used as dependent data. Country data from 1999-2002 and instability indicator data from 2003-2006 was held out of the model to test the predictive ability of the model. The result was a 150 x 4 canonical correlation scores matrix which was used to create an OLS regression model for each stability indicator. Part of these scores can be found in Appendix L.

When the canonical correlation scores of the training data was fitted via OLS to undernourishment data with a 4 year lag, the following results and model parameters were obtained:

Regression Statistics	
R-square	0.9267
R-square adj	0.9247
RMSE	4.1405
PRESS	2657.8
R-square Predicted	0.9217

	df	*SS*	*MS*	*F Ratio*	*Prob > F*
Regression	4	31451.1	7862.8	458.6	<.0001
Residual	145	2485.9	17.14		
Total	149	33937			

	Coefficients	*Standard Error*	*t-Stat*	*P-value*
Intercept	38.636	0.3381	114.28	<.0001
Canonical Variate 1	11.678	0.3392	34.43	<.0001
Canonical Variate 2	-8.44	0.3392	-24.88	<.0001
Canonical Variate 3	-1.858	0.3392	-5.48	<.0001
Canonical Variate 4	0.1055	0.3392	0.31	0.7562

Table 4-3. Parameters of Undernourishment Model

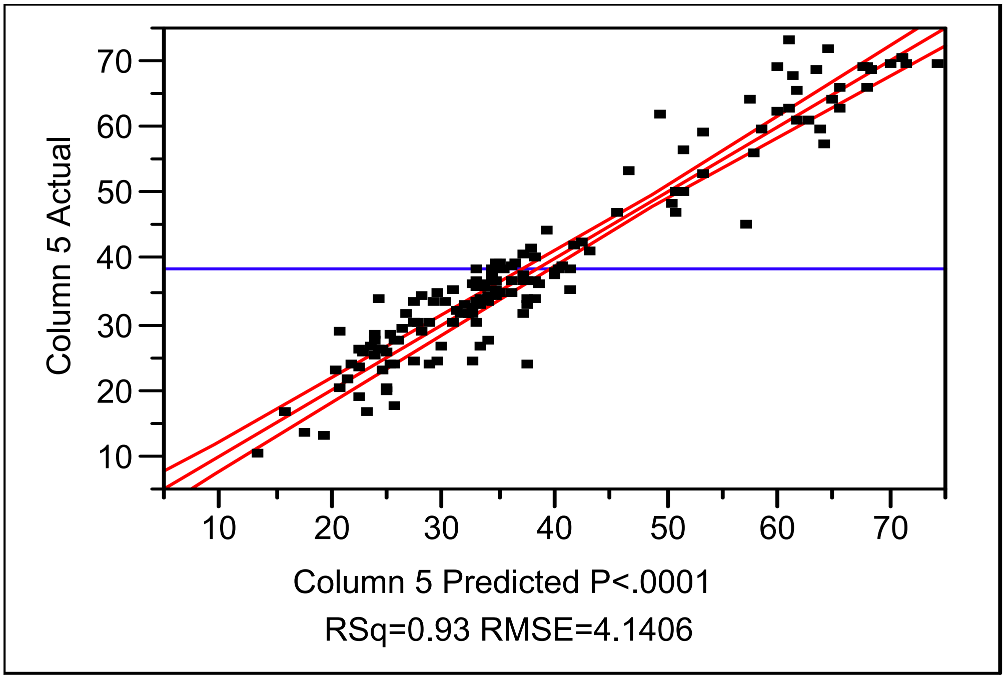


Figure 4-2. JMP Actual vs. Predicted Values of Training Data in Undernourishment Model

The equation describing the model in Table 4-2 would thus be:

$$\hat{y} = \beta_0 + \beta_1\omega_1 + \beta_2\omega_2 + \beta_3\omega_3$$

where

β_i are the regression coefficients, in this case for MI 1 β_i are 38.64, 11.68, -8.44, and -1.858 and ω_i are the canonical variate scores

The model in Table 4-2 displays both low multi-collinearity, and near constant variance. The Variance Inflation Factor (VIF) for each of the four canonical scores is only .0067, where a score greater than 5 or 10 is required to indicate multi-collinearity (Montgomery, *et al*, 2006: 307). The variance of the residuals in Figure 4-2 appears near constant.

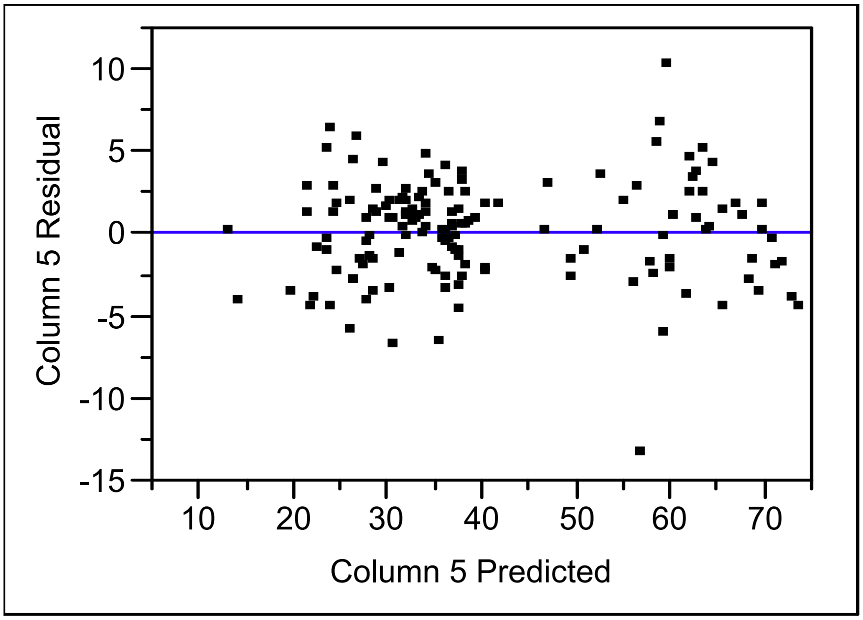


Figure 4-4. Undernourishment Model Residual Using JMP

The outlier point represents the prediction of malnutrition in Ethiopia in 1991. This outlier is an artifact of the imputation process, where the imputed undernourishment dependent data point lies at the low end of the expected range. The undernourishment data points for the year prior, and the year after are at least 15% higher than the 1991 data point. The model is therefore an excellent fit for the training data.

The model shown above was tested on the hold out data set with good results. Table 4-4 on the following page shows the by country and year results of the model on the hold out data, as well as the overall performance of the hold out model in terms of R^2 and RMSE.

		Malnutrition		
Country	Year	Actual	Predicted	Error
Yemen	2003	37	34.3166	-2.6834
Yemen	2004	38	38.1005	0.1005
Yemen	2005	30.4838	36.092	5.6082
Yemen	2006	36.6368	34.3199	-2.3169
Somalia	2003	16.1	23.5156	7.4156
Somalia	2004	19.7	19.6933	-0.0067
Somalia	2005	12.7	19.0525	6.3525
Somalia	2006	18.6	16.1868	-2.4132
Sudan	2003	27	29.0543	2.0543
Sudan	2004	26	31.8124	5.8124
Sudan	2005	29.2417	30.5496	1.3079
Sudan	2006	32.3014	28.9256	-3.3758
Djibouti	2003	26	24.7069	-1.2931
Djibouti	2004	24	22.5199	-1.4801
Djibouti	2005	25.2437	18.2103	-7.0334
Djibouti	2006	27.5347	15.1488	-12.3859
Kenya	2003	31	33.4802	2.4802
Kenya	2004	31	29.5709	-1.4291
Kenya	2005	28.2095	30.6891	2.4796
Kenya	2006	27.6009	33.7472	6.1463
Ethiopia	2003	46	40.94	-5.06
Ethiopia	2004	46	42.3055	-3.6945
Ethiopia	2005	48.6663	44.5684	-4.0979
Ethiopia	2006	48.366	44.1047	-4.2613
Eritrea	2003	73	64.3102	-8.6898
Eritrea	2004	75	68.7003	-6.2997
Eritrea	2005	72.2599	71.0569	-1.203
Eritrea	2006	71.7162	71.7176	0.0014
Rsquared		0.9152		
RMSE		4.9846		

Table 4-4. Hold Out Data Results for Undernourishment Model

The R^2 and the RMSE of the hold out data is larger than that of the training data, but that is to be expected. The difference is not great, and the model

$\hat{y} = \beta_0 + \beta_1\omega_1 + \beta_2\omega_2 + \beta_3\omega_3$ is recommended as a forecasting model of undernourishment based on the low RMSE, high R^2, and apparent robustness of the model as demonstrated by its application to the hold out data.

By examining the results of Tables 4-1 and 4-3, the direction of effect of each raw variable can be seen. For instance, the 4 year time lagged existing level undernourishment is negatively weighted on the independent canonical correlation variate 1. Canonical correlation variate 1 has a negative β coefficient in the OLS model, meaning that as current malnutrition goes up, canonical score 1 decreases, and thus the predicted level of undernourishment four years in the future goes up. Only one of the variables had the opposite effect on the model that is predicted by the experts. The surprising variable was trade openness. The more trade openness, the higher the predicted malnutrition. This does follow the pattern of the data, though, where Eritrea has higher trade openness scores than most countries, while having the worst levels of undernourishment of all the countries examined in this study. A possible explanation is that trade openness, as it is defined in Appendix C, represents self sufficiency. It could also mean that the food grown in a country is being sold abroad, rather than at home, leading to shortages. Interpreting the reason why greater trade openness causes the predicted level of undernourishment to rise is not easily deciphered, and is an area for future research.

In summary, a continuous model of undernourishment in the HoA region using canonical correlation scores based on the 54 independent variables listed in Appendix K, and the 4 types of instability indicators to build an OLS model of the dependent variable undernourishment is recommended. The exact parameters of the model will change slightly depending on the imputed data. The advantages of this model are it's adherence to the assumption of OLS, its accuracy both in terms of R^2 and variance, and its simplicity. The raw variables with the greatest effect on the model are the existing levels of refugees, battle deaths, caloric intake, and undernourishment. Additionally, trade openness is strongly weighted on the second variable in the model. Given the correlation between undernourishment and the other three instability indicators, this model may provide an analyst with a "big picture" forecast of the strength and stability of a regional government, as well as of simple undernourishment.

4.3.2. Discrete Models for Indicators of Instability

The overarching goal of this study was to produce predictive models of instability indicators Horn of Africa that provide forecasts for each individual country. While an accurate continuous model might provide a more specific estimate with a prediction interval, rather than a simple forecast of the observation being above or below a pre-determined threshold and an associated percent probability (in the case of logistic regression). However, given the high variance of the continuous models and the widespread acceptance of discrete models such as PITF and Collier and Hoeffler, the study also attempted to improve on the accuracy of existing discrete models which make categorical forecasts. The results of those studies were discussed briefly in Chapter 3. The results of this study are compared with the PITF results since the PITF scores represent the smallest hold out data apparent error rates of previous studies, and this study shares some dependent data with the PITF dependent data set.

This section describes the data in this section of the study and its relationship to Durch's hypothesis. It presents models for each of the stability indicators except undernourishment, for which an accurate continuous model has been developed and described in section 4.3.1, and discusses which variables are most influential for each of the preferred models. The data is examined for trends. A best model is recommended based on Apparent Error Rate (APER) and false negative predictions.

4.3.2.1 Discrete Models and Durch's Hypothesis

In the literature review for this study, nowhere could be found an agreed upon definition of exactly hold many refugees, battle deaths, or genocide / politicide deaths constitutes an instability event. Thus, this study examined a spectrum of each as thresholds with which to discriminate between observations. The battle deaths, refugees, and genocide dependent data was examined by first arranging all the data points in ascending order. The initial results supported Durch's concept that countries could continue to function for a long period of time, despite inherent conditions promoting destabilization. However, a trigger event would also disrupt this quasi-stability, resulting in catastrophic destabilization. The plots are shown in Figures 4-4, 4-5, and 4-6. Each data point in the graphs represents the number of battle deaths, refugees, or genocide deaths in one country in one year. They are arranged from smallest (0) on the left side of the graph to the largest observation on the right. There area total of 206 data point for each of the three instability indicators.

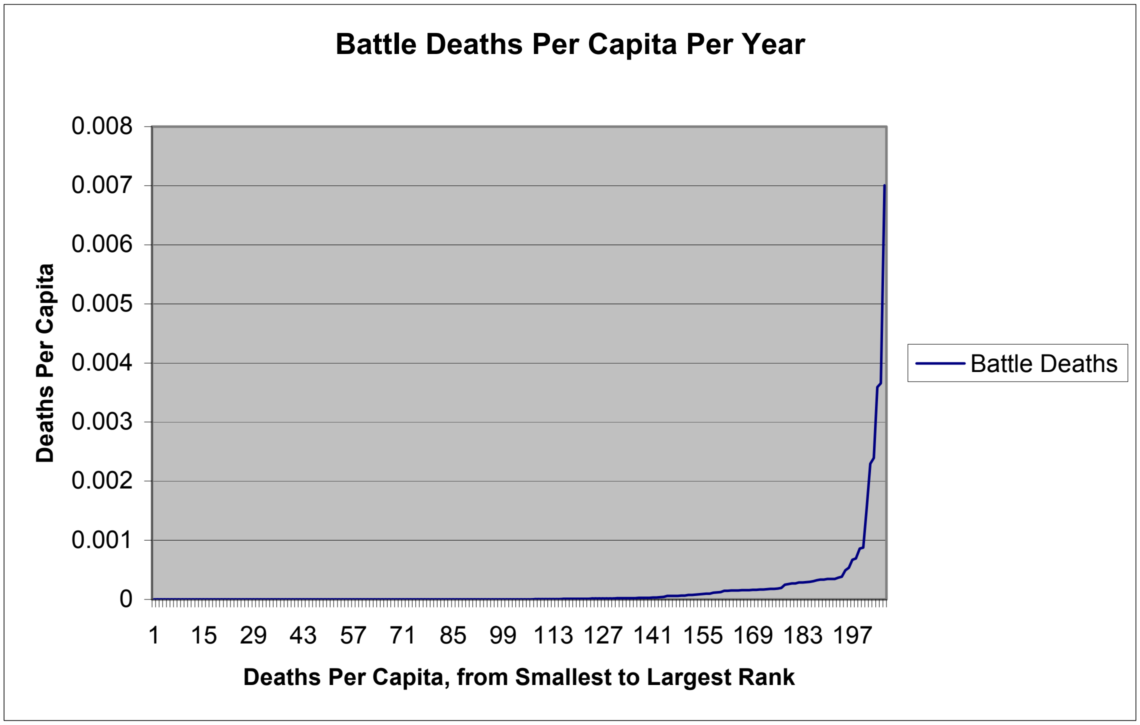


Figure 4-4. Battle Deaths Per Capita in Ascending Order

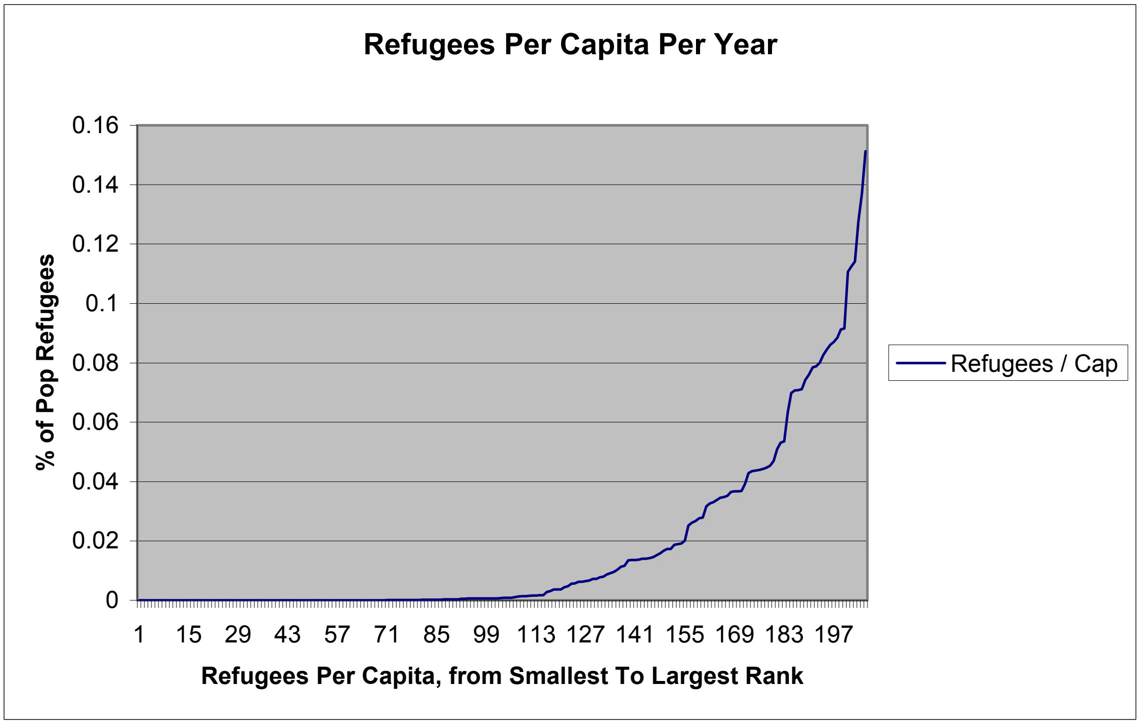


Figure 4-5. Refugees By Year in Ascending Order

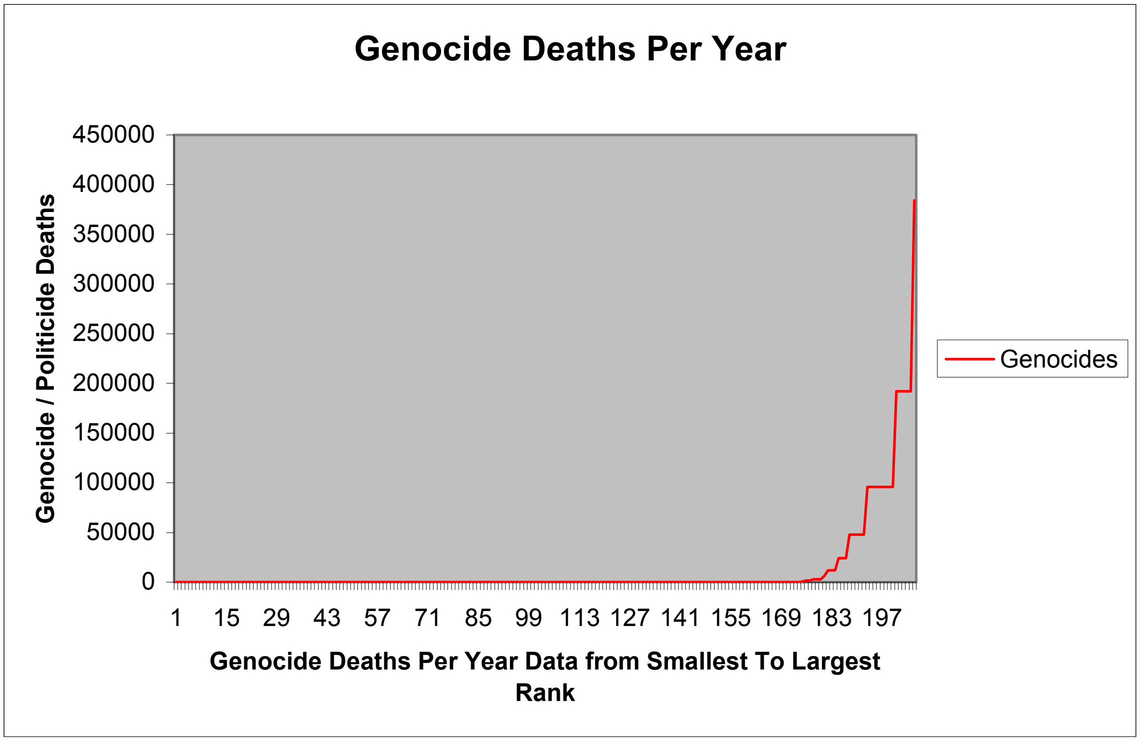


Figure 4-6. Genocide / Politicide Deaths By Year in Ascending Order

Based on Durch's model, as well as the CAA's "oily rag" concept, there should be a dividing point between countries which are at least marginally stable, and those countries which have destabilized, since the slope of Durch's line, and those shown above rapidly increases out of the stable (flat or nearly flat) range. The idea is that countries do not exist for very long in the transition state between stability and instability, and thus can be classified into one of the two categories, whether stability is defined by battle deaths, refugees, or genocide. Hence the graphs should exhibit a narrow band where the slope of the line is changing rapidly upward. The graphs were used to find the area where this transition begins to occur, and then finding the number of battle deaths per capita, refugees per capita, or genocide deaths on the y-axis associated with this point on the line. The inflection on each line suggests a threshold for classifying countries using discriminant analysis (DA) and logistic regression.

Polynomial curves were fitted to each of the graphs in Figures 4-4 through 4-6, minus the tails where the data is zero. The second derivative was taken of each and the roots of the resulting polynomial found using MATLAB. The largest solution was taken as the point at which the curve begins become undamped. For battle deaths, this occurred at about the 193rd observation, which equates to about 1250 battle deaths. This suggests that the most accurate threshold for predicting battle deaths will be near 1200. However, it should be noted that the threshold for intervention, which this study is concerned with, is almost certainly far less than this. A conflict resulting in this level of violence would require a more complex and riskier response by the international community. In the past decade, there has been no direct intervention in Eritrea, Ethiopia, Sudan, or most recently Kenya despite casualties both higher and lower than this threshold. Thus, the thresholds explored for battle deaths in this study go from 0 to 1250. As will be seen later, the best threshold point in this range is well below 1250.

In the cases of refugees and genocides the expected best performance threshold based on the graphs was estimated as approximately 250,000 and 0 respectively. The figure for refugees poses the same problem the threshold of 1200 did for battle deaths. 250,000 refugees is a massive humanitarian crisis, and a number far smaller would still be indicative that a government has failed to provide of a large segment of its populace. For the purpose of this study, refugee thresholds smaller than 250,000 total per country will be examined in greater detail. In addition, the actual analysis did not support 250,000 as an optimal threshold. PITF used 0 genocide deaths as a threshold. Other genocide thresholds besides the PITF's are explored in this section however. Additionally, using per capita data for each instability indicator were explored in the same manner as the raw data, and actually provided better models for battle deaths and refugees.

4.3.3. Discrete Models of Battle Deaths

Initial pilot studies using both discriminant analysis and logistic regression were conducted using 39 variables with no transformations, such as normalizing the data or converting it to PCA or canonical correlation scores. The pilot studies were done for each of the three remaining instability indicators (battle deaths, refugees, and genocide). These 39 variables were chosen based on the conditioning of the matrices they developed for both pooled covariance matrix and the inverse of the data within the logistic regression function in MATLAB. Variables that were linear combinations of other variables in the model were identified using the pooled covariance matrix, its inverse, and the inverse of the data matrix. The variables used in these pilot studies are listed below.

Year	Religious Fractionalization
Literacy	Linguistic Fractionalization
Gender Parity	Transition Government
Primary Commodity Exports	Full Autocracy
Life Expectancy	Partial Autocracy
Infant Mortality Rate	Partial Democracy w/Factionalism
Youth Bulge	Political Discrimination
Trade Openess	Economic Discrimination
Urban Population	Years since last conflict
Telephone Subscribers per 100	Change in Calories
Aid as a % of GNI	Change in Infant Mortality Rate
Military as a % GDP	Pct Paved Roads
Agriculture as a % GDP	Calories Per Day Per Capita
Durability	Education as a % GNI
Trade Ratio	Water Per Capita
Aid per Capita	Population Density
GDP Growth	Arable Land Per cap
Missing Data	Km Road per Cap
Bad Neighbors	Relative GDP Per Cap
Ethnic Fractionalization	

Table 4-5. 39 Variables Used In Discrete Forecast Pilot Studies

For battle deaths, the threshold for classifying a country as being in a state of “hot war”, and thus in a position which limits non-military options, was tested at all battle deaths between 0 and 1250, giving a total of 175,000 hold out observations, based on 28 hold out observations tested at 1250 thresholds. The reason 1250 was selected was that only 2 observations in the hold out set exceed 1250; it was assumed that 1250 indicates a war where intervention would be difficult and the international community reluctant to become involved in an armed conflict. An example of this is the U.N.’s complete withdrawal from Iraq after its headquarters was attacked. The U.N. still has not returned (CNN, 2003). Appendix S shows the results of the models which used logistic regression and DA. The initial results for models of battle deaths were promising, but still not as accurate as previous attempts by PITF, albeit on a 4 years forecast vice a 2 year. The initial pilot studies had average error rates for battle deaths as low as 25%.

Because the distributions of the raw data were rarely normal, and thus in violation of one of DA's assumptions, the study was repeated with some of the data transformed (where possible) to make the data more normally distributed, or to at least reduce skewness. The lack of improvement in PAER was not unexpected, since non-normality with discriminant analysis primarily biases the error, rather than the overall prediction error rates (Dillon, 1984: 363). These transformations are listed in Appendix I. However, transforming the data did not produce a superior model for battle deaths, or any of the other instability indicators.

As a follow on, raw data was replaced with 13 PCA scores and also 4 canonical correlation scores and used in separate pilot studies of each of the three remaining instability indicators. The data used to generate these scores were the same 54 variables used to develop continuous models in section 4.2, and are listed in Apendix K. The canonical correlation scores had a higher apparent error rate than similar models using raw data or PCA scores, thus the results are not shown in Appendix S. This is not surprising, since DA and logistic regression use distances and variance to discriminate between groups, while canonical correlation scores capture correlation. The use of PCA scores in the model demonstrated marked improvement in terms of battle death apparent error rate, or the number of times the model incorrectly classifies a hold out observation divided by the total number of hold out observations. An incorrect classification occurs when an observation is predicted to be within one group, defined by having more or less than the threshold value, when it should actually belong to the other group as defined by the same threshold. In the case of using 54 variable converted to 13 PCA scores with total battle deaths per year by country, the error rate was 19.2% using DA, and 21.4% using logistic regression to discriminate between the classes.

A further attempt at obtaining a better Apparent Error Rate (APER) was made by using battle deaths per capita instead of the simple raw total by country by year. This was done to scale the stability indicator based on nation size, to enhance the comparability of the models, and allow the use of the same analytic program. A country with a population of 2 million losing 1000 people in a war would seem to be more likely to experience destabilization than a country of 300 million experiencing the same number of casualties, all other things being equal. The use of per capita battle deaths in turn yielded the most accurate forecasting model of battle deaths: using logistic regression on varimax rotated PCA scores of 54 variables to forecast battle deaths per capita resulted in an APER on the hold out data across the five multiply imputed data sets of .1706, or a pseudo R^2 of .8294. The next closest APER was obtained using the same PCA scores, DA, and raw battle death data resulting in an APER of .1929. When a 95% confidence interval is applied to the APERs of the logistic regression using PCA scores and per capita data model the APER becomes .1706 +/- .052, placing the next closest model's mean within the interval. Assuming a normal distribution, this still indicates there is an 88.1% chance the true mean of the best observed model of battle deaths APER is less than the observed mean of the second best model.

Some of the difference between models using raw data vice per capita data comes from the difference in data structure caused by deaths being per capita, rather than a raw integer. There as many observations in the data set with more than .000125 battle deaths per capita as there are with more than 2000 total battle deaths in a year, and vice versa. This was done deliberately. These values both approximately correspond with the 160th smallest observation of battle death and battle deaths per capita. If one assumes that 0 battle deaths per year and 0 battle deaths per capita are the best case scenarios, and that the largest observed values of battle deaths and battle deaths per capita are the worst case scenario, selecting a maximum threshold that corresponds with the same point on both of the ordered scales ensures the models for each type of dependent data are examining the same general spectrum of outcomes. Additionally, having a threshold too high can result in training matrices that are more likely to have fatal singularities.

At this point, it was decided to model battle deaths per capita based on using 54 variables converted to 13 rotated PCA scores, and using logistic regression to discriminate between those observations where the number of battle deaths exceeds the set threshold. This decision was based on the battle deaths per capita model described above having the lowest observed APER, and a bias towards false negatives over false positives. Thus, further discussion of battle deaths per capita will focus on this model framework and its results.

The best case model showed a bias towards false positive errors, with 68.6% of the errors being false positives and false positives indicating that the battle deaths per capita model predicted a country would be above the threshold for a given year, when in fact it actually had less battle deaths per capita than the threshold in that year. This was deemed to be more desirable than the converse, the false positive. This is due to the assumption that it is cheaper to have made efforts to improve a nation that was not going to have fallen into a state of conflict, than to sit by and do nothing and clean up the consequences of a war later. The MATLAB program records the errors every time the predictions change, and keeps a tally of the errors. A prediction of battle deaths per capita exceeding the threshold was scored a 1, and a prediction of battle deaths below the threshold a 0. In the error reports, false negatives are scored as a -1, and false positives are scored as a 1. Total prediction errors, analogous to area under the curve in Figure 4.7, are shown in Appendix S. The table below shows the total errors by country using the change log, and how many prediction changes occurred, along with a percentage score showing what types of errors each prediction typically had. The change log recorded the predictions of the model as 1's and 0's for each hold out data point, and the errors of these predictions, every time the prediction changed due to the shift in the threshold as it was iterated.

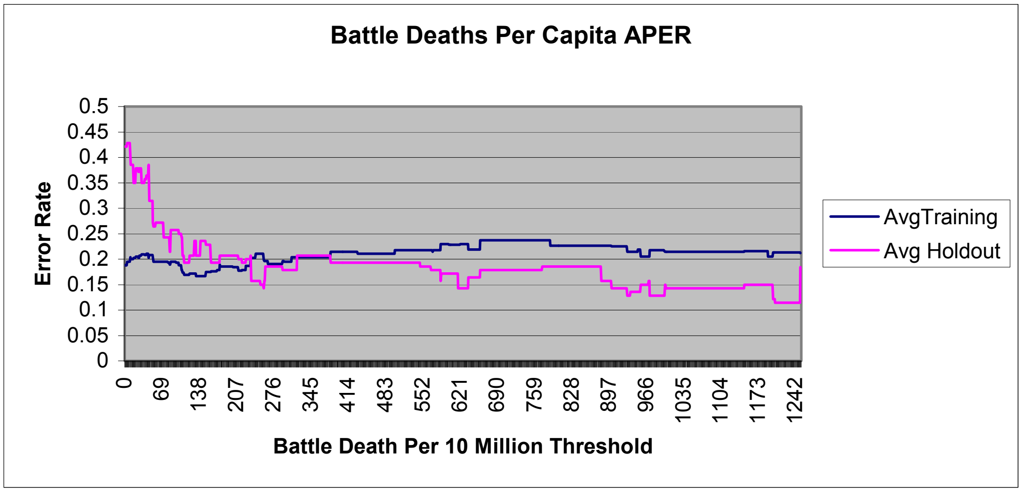


Figure 4-7. Battle Death APER Using Logit Regression and Per Capita Data

The Figure 4-7 above uses aggregated error data from all 5 multiply imputed data sets. It shows the model performs poorly at the lowest thresholds of zero, where there is less to separate countries with violence that is "noise", or is simply the inherent level of violence within the nation that is insufficient to cause full scale war or destabilization. However, the accuracy of the predictions, both for the training and the hold out data, stabilizes above approximately 100 battle deaths per 10 million population. This graph demonstrates the model's robustness at most threshold settings, and allows subject matter experts to select a threshold for instability events and have confidence the model will provide accurate predictions.

		Actual	
		Over	Under
Predicted	Over	11209	20501
	Under	9361	133929

Table 4-6. Recommended Battle Deaths Per Capita Model Confusion Matrix

The confusion matrix in Table 4-6 shows the aggregated model predictions versus the actual outcomes of the hold out data using the spectrum of thresholds. It shows how many correct predictions of each type, and incorrect predictions were made using the spectrum of threshold tested on each of the 5 data sets, which were then tallied. "Total Changes" in Table 4-7 indicates how many times the predictions changed as the threshold was incremented. Error trend indicates the number of false positives less the number of false negatives. % Error indicates the error trend divided by the total number of changes. Thus, when predicting Ethiopia's battle deaths per capita in 2003, the model overestimated the expected amount of violence in at least 20 out of 36 thresholds.

Country	Year	Error Trend	% Error
Yemen	2003	0	0.000
Yemen	2004	0	0.000
Yemen	2005	0	0.000
Yemen	2006	0	0.000
Somalia	2003	8	0.222
Somalia	2004	8	0.222
Somalia	2005	0	0.000
Somalia	2006	-30	-0.833
Sudan	2003	-1	-0.028
Sudan	2004	-7	-0.194
Sudan	2005	6	0.167
Sudan	2006	4	0.111
Djibouti	2003	7	0.194
Djibouti	2004	7	0.194
Djibouti	2005	7	0.194
Djibouti	2006	7	0.194
Kenya	2003	-4	-0.111
Kenya	2004	-3	-0.083
Kenya	2005	-9	-0.250
Kenya	2006	-17	-0.472
Ethiopia	2003	20	0.556
Ethiopia	2004	-4	-0.111
Ethiopia	2005	-1	-0.028
Ethiopia	2006	8	0.222
Eritrea	2003	21	0.583
Eritrea	2004	23	0.639
Eritrea	2005	19	0.528
Eritrea	2006	3	0.083
Total Changes		36	

Table 4-7. Battle Death Prediction Errors

The logistic regression model appears to consistently under predict Somalia's violence, particularly the fighting between the Western backed provisional government and the Islamic Court's forces in 2006 (BBC News, 2006). This can bee seen in Table 4-

6, where the model underpredicted Somalia's battle deaths per capita in at least 30 out of 36 cases, with an error trend of -30. It consistently over estimated the amount of violence expected in Ethiopia in 2003, and it consistently over predicted the number of battle deaths per capita in Eritrea for each hold out year except 2006.

It should be noted that the model is consistently under predicting Kenya's battle deaths per capita at an increasing rate. These mis-classifications are a result of historical "inertia". To state it simply, Eritrea has had the highest levels of battle deaths per capita in the region. Thus, the model expects that countries that "look" like Eritrea should expect high levels of violence. Conversely, Kenya has traditionally had very low levels of battle deaths and a relatively stable government, thus the model tends to predict countries that look like Kenya are less likely to experience violence. Given more data as well as more outbreaks of violence, the model might be able to "recognize" more indicators of future conflict. However, to more quickly identify when a country's fundamental nature has shifted, some form of turning point analysis would seem appropriate. This is discussed in Chapter 5.

Figure 4-7 suggests the model may not perform as well against holdout data with a low threshold near zero. At the threshold of zero the .44 APER of the hold out data approaches the accuracy of a coin flip. This is undesirable, since previous research by PITF set the threshold at 10 total fatalities (e-mail correspondence with Dr. Ulfelder). The optimal or subject matter expert preferred threshold may well be above this number. It should also be noted that there have been many skirmishes in the past 50 years with more than 10 battle deaths which have not led to complete destabilization, notably between North Korea and South Korea, and US and Soviet aircraft. However, the battle deaths threshold is also almost certainly well below 1250 or 2000 per year per country or their per capita equivalents.

The secondary purpose of this study was to identify which factors are most significant in each model, in this case battle deaths per capita using PCA scores and logistic regression. The study verified that rotating the factors did not alter the predictions obtained using the resulting PCA scores, for battle deaths per capita or the other instability indicators. For logistic regression, the apparent error rate in the training data stabilizes above 936 battle deaths per 10,000,000, as well as having the lowest hold out data and training data APER. Thus, setting .000093 as the threshold and using standardized inputs from Multiply Imputed data set 2 (MI 2), Table 4-7 summarizes the T-scores and P-values found using a logistic regression model with PCA scores obtained from a varimax rotated loadings matrix and per capita battle deaths as a dependent variable. The test and *p*-values for each rotated principal component score is shown in Table 4-7. The loadings on the rotated principal components are shown in Appendix T.

	0 Threshold			100 Threshold			936 Threshold		
	Coefficient	t-test	p-value	Coefficient	t-test	p-value	Coefficient	t-test	p-value
Intercept	31.1098	1.7847	0.0743	10.2813	0.4945	0.621	-24.5692	-1.1817	0.2373
Wealth and Urbanization	0.2126	2.2873	0.0222	0.1768	1.8702	0.0615	0.0046	0.0443	0.9646
Fractionalization	-0.1994	-1.6003	0.1095	-0.0443	-0.2768	0.7819	-0.2017	-1.0149	0.3101
Agricultural Reliance / Stress	-0.2243	-1.7394	0.082	-0.3627	-2.8099	0.005	-0.1856	-2.3293	0.0198
Total Population, Ethiopia	-0.2666	-1.4738	0.1405	-0.0325	-0.2562	0.7978	-0.3519	-2.5783	0.0099
Time Effects	0.239	1.7041	0.0884	0.1031	0.6722	0.5014	-0.0469	-0.2912	0.7709
Military / Health Spending	-0.3469	-2.6559	0.0079	-0.2801	-2.189	0.0286	-0.158	-1.2183	0.2231
Economic Integration into Global Market	0.8016	3.6815	0.0002	0.8006	3.7543	0.0002	0.4565	2.6696	0.0076
Long Term Government Stability	-0.1507	-0.8858	0.3757	-0.4158	-2.2095	0.0271	-0.4315	-1.8437	0.0652
Weak Government / Existing Instability	-0.2379	-1.0061	0.3144	0.0874	0.4735	0.6359	0.1792	1.1224	0.2617
Change in Infant Mortality Rate	-1.2274	-3.0298	0.0024	-1.3277	-3.0765	0.0021	0.0271	0.1504	0.8805
Political Discrimination / Exclusion	-0.218	-0.86	0.3898	-0.1568	-0.494	0.6213	0.1347	0.4476	0.6544
Improving Quality of Life	0.1837	0.8757	0.3812	0.3553	1.6257	0.104	-0.1732	-0.9536	0.3403
Transition Government	-0.0182	-0.0743	0.9408	-0.0669	-0.3115	0.7555	-0.0214	-0.1114	0.9113

Table 4-8. *T*-tests and Associated *p*-values for Battle Deaths Logistic Regression Model

Larger magnitude *t*-test values indicate a variable is more significant to the logistic regression equation at a particular threshold. The *p*-value is based on the *t*-value and the number of degrees of freedom, and is indicative of the probability that the variable does not contribute significantly to the logistic regression model. Note the difference between the models at different thresholds, indicating the underlying structure of the model changes depending on threshold setting. Given the desire for accuracy at lower thresholds, since low level violence typically precedes more intense conflict, as well as the implicit assumption that the variables contributing to accurate models provide greater insight, the data associated with a threshold of 100 battle deaths per 10 million population is discussed here. Additionally, this threshold is used in the model comparison section as well. The proposed meanings of each of the principal components was shown in Table 4-2. Table 4-9 above suggests that economic integration into the global market (PC 7) is important to the battle deaths per capita model, no matter the threshold. At lower thresholds change in infant mortality rate (PC 10) may be used as an indicator of impending conflict. Reliance on agriculture, and scarcity of agricultural resources (PC 3) is a significant factor no matter where the threshold is set for the region. While long term government stability (PC 8) does not significantly contribute to the model when the threshold is set to 0, it is important in the HoA region at higher settings. This suggests that while a stable government does not prevent smaller skirmishes, it may act as a brake that prevents the situation from spinning out of control. Overall the results in Table 4-7 indicate that trade ratio and change in infant mortality may be the most important individual variables to monitor, since these two variables are the most highly loaded variables on two of the most significant principal components in the model.

In conclusion to section 4.3.2, the recommended model for battle deaths per capita is using PCA scores and logistic regression. The model demonstrates robustness over most of the threshold range, and provides an APER over the tested range of only .1706, while being biased towards the less harmful false positive error type. This model has the advantage of being reproducible with readily available statistical packages, and can be implemented to anticipate which countries are more likely to be in a state of conflict 4 years from now. Additionally, the observation that the variables change in infant mortality rate and trade ratio are highly weighted on the two of the significant principal components suggests that if one were to look for early warnings of instability these two might provide initial indications. They also have the advantage that they are both consistently tracked in the region, except for the latter in Somalia.

4.3.4. Discrete Model of Refugees

The next discrete model of an instability indicator was developed for refugees. The same modeling approaches that were analyzed for battle deaths were used on the refugee data. This included the pilot study data, DA and logistic regressions, raw data, rotated PCA scores, refugees per capita dependent data, and raw dependent refugee data. The results of each of these models are shown in Appendix S. Using logistic regression as a discriminating method, there was no statistical difference between the error rates of models using raw refugee data and per capita data to define the threshold. However, using per capita refugee data produced a significantly lower number of false negatives. The per capita data model also showed a lower level of variance in the APERs than using raw data. These two factors lead to the choice of using PCA scores as the independent variables and per capita data with logistic regression over using raw refugee data with

PCA scores and logistic regression. Using the same 13 rotated principal components from the battle deaths models and logistic regression to forecast refugees per capita produced a model whose aggregated hold out data APER across all 5 data sets was .0588. The training data APER was .08329 predicting refugees per capita using the optimal model described in the previous sentence. The difference between the hold out APER and the training data APER shows a positive bias for the holdout data to have a lower APER since the 95% CI on the mean of the hold out data APER was .0588 +/- .0144. Additionally this model has relatively low variance, with 95% confidence upper and lower bounds of 1.12e-5 and 1.429e-4 which places the variance of the second best model barely within the range (1.242e-4), indicating that there is a strong probability the variance of the preferred model is smaller. (Wackerly, *et al*, 2002: 407). It has been suggested in the past that the bias on hold out data could represent a structural shift over time such as the fall of the Soviet Union. PITF investigated the hypothesis that the fall of the Soviet Union caused a structural change, but eventually rejected it (Goldstone, *et al*, 2005: 15). Further references to the model of refugees per capita or the refugee model refer to using 54 variables to produce 13 rotated principal components and logistic regression to forecast refugees per capita.

The range of thresholds chosen selected for the raw refugee data was from 0 to 100,000, with thresholds between 0 and 100 incremented by 1, incremented by 10 between 100 and 1000, and incremented by 100 between 1000 and 10000, and by 1000 between 10000 and 100000. Refugee per capita thresholds were incremented similarly on a semi-logarithmic scale, starting at 0 and ending at 7200 refugees per million. The logarithmic scale was chosen to reduce the total number of thresholds explored by the program written to compile results of the models (shown in Appendix J), and because the granularity of the refugee data increased as the number of refugees rose, i.e. when refugees were above 10000, the UN data was typically rounded to the nearest 1000, making observations based on smaller increments redundant. This selection was also based on the observed distribution of data seen in Figure 4-5, and a subjective assessment that 100000 refugees represents a point at which a instability has taken place and the world press has already reported the instability. The latter figure was chosen since 100000 refugees represented the 130th smallest refugee observation, and 7000 per million represented the 130th smallest refugee per capita observation. This was done to ensure that the upper threshold of refugees per capita represented a similar level of severity to that of raw refugee data range, thus making the results more comparable. The Figure 4-8 shows the aggregated from all five data sets using rotated PCA scores classified by logistic regression to build a model of refugees per capita.

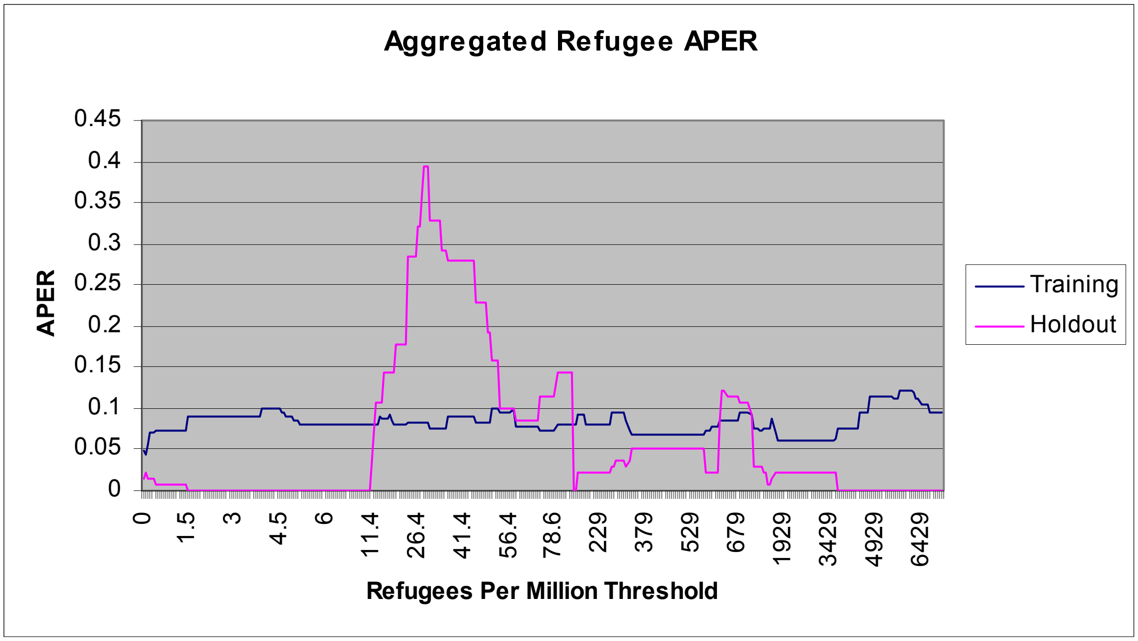


Figure 4-8. Aggregated APER for PCA Scores with Logistic Regression

Refugees Per Million	**0**			**450**			**7200**		
Test	Coefficient	t-test	p-value	Coefficient	t-test	p-value	Coefficient	t-test	p-value
Intercept	223.8	1.5924	0.1113	226.5719	2.8508	0.0044	172.5062	3.2618	0.0011
Wealth and Urbanization	0.0624	0.1938	0.8463	0.3687	2.4242	0.0153	0.1216	0.3292	0.742
Fractionalization	2.5696	3.8588	0.0001	1.3209	1.7175	0.0859	1.5662	1.3867	0.1655
Agricultural Reliance / Stress	-2.4411	-1.6964	0.0898	-1.549	-3.6002	0.0003	-1.238	-2.7428	0.0061
Total Population, Ethiopia	-7.1978	-2.8687	0.0041	-0.6867	-1.7403	0.0818	0.943	1.3209	0.1865
Time Effects	2.7219	3.3958	0.0007	2.12	4.0941	0	1.1982	3.391	0.0007
Military / Health Spending	-2.3516	-3.894	0.0001	-0.5847	1.3149	0.1886	0.7863	0.9671	0.3335
Economic Integration into Global Market	0.73	0.9874	0.3234	0.461	2.4002	0.0164	1.9996	2.5318	0.0113
Long Term Government Stability	0.7475	0.4147	0.6783	-0.873	-2.7715	0.0056	-0.0173	-0.0444	0.9646
Weak Government / Existing Instability	0.5697	0.2083	0.835	3.3476	3.7498	0.0002	2.0124	2.5947	0.0095
Change in Infant Mortality Rate	-0.7659	-0.6184	0.5363	-2.8908	-1.965	0.0494	-1.1515	-1.8516	0.0641
Political Discrimination / Exclusion	-2.5428	-1.6268	0.1038	-1.0135	0.2769	0.7819	-0.1214	-0.2137	0.8308
Improving Quality of Life	1.8704	3.4008	0.0007	1.0879	1.2726	0.2032	0.3556	1.2563	0.209
Transition Government	-0.6897	-0.1491	0.8815	-0.3304	-0.4941	0.6212	-1.1419	-1.3269	0.1845

Table 4-9. PC Significance Test Results for Refugees Per Capita Model

The graph in Figure 4-8 shows the aggregated error on hold out data from analysis done with all five multiply imputed data sets. This shows the threshold regions in which the model functions well, and in which it has a higher error rate. The bulge from approximately 11.4 to 78.6 represents a region in which the parameters which define the model are changing. Referring to Table 4-9 it can be seen that the variables which are significant to the model change between 0 and 450, and the “bulge” in error rate exists in the area where this change is taking place. This suggests a three group discriminant function might be appropriate to modeling refugees per capita. Keeping in mind the pseudo-logarithmic scale of this graph, it can be seen that much like the model of battle deaths per capita the refugees per capita model also has it’s worst prediction rate at the low end of the spectrum. This also seems to indicate that some amount of refugees is endemic to countries in the region, and there is little structural difference between a countries with refugees in the range in the bulge. This makes the region noisy, and reduces the forecast accuracy.

The types of errors exhibited by the model are shown in figure Table 4-9. Each time the model's errors changed as the threshold moved, the new set of errors was recorded and tallied. The confusion matrix for all five data sets is shown in Table 4-10. Table 4-11 shows the same type information, and was collected in the same way as Table 4-7.

		Actual	
		Over	Under
Predicted	Over	34428	2447
	Under	432	14213

Table 4-10. Recommended Refugees Per Capita Model Confusion Matrix

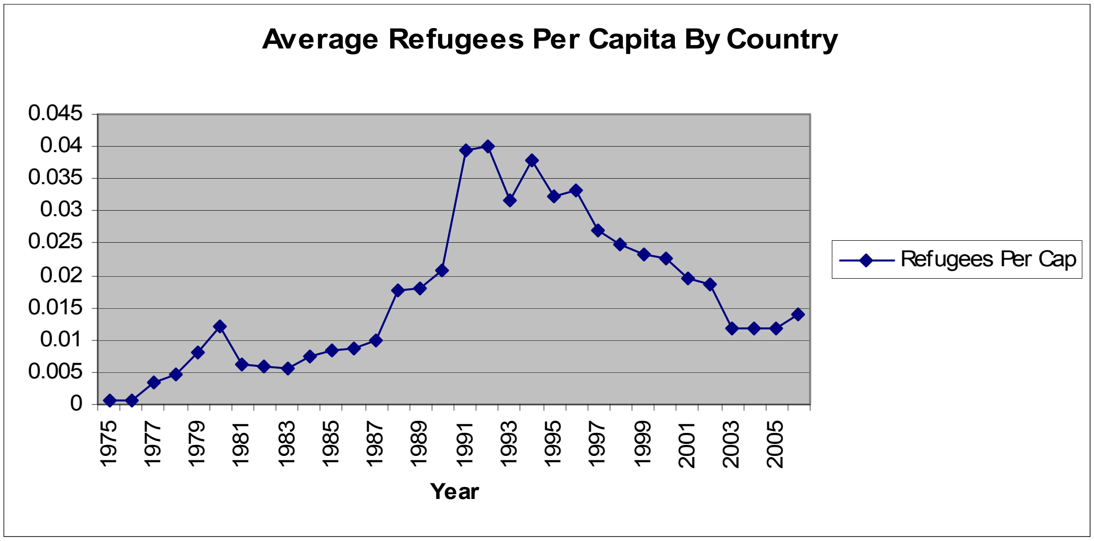
Country	Year	Total Log Trend
Yemen	1999	6
Yemen	2000	8
Yemen	2001	11
Yemen	2002	9
Somalia	1999	0
Somalia	2000	0
Somalia	2001	0
Somalia	2002	0
Sudan	1999	0
Sudan	2000	0
Sudan	2001	0
Sudan	2002	0
Djibouti	1999	9
Djibouti	2000	11
Djibouti	2001	13
Djibouti	2002	11
Kenya	1999	-2
Kenya	2000	2
Kenya	2001	-6
Kenya	2002	-8
Ethiopia	1999	4
Ethiopia	2000	3
Ethiopia	2001	-2
Ethiopia	2002	-4
Eritrea	1999	0
Eritrea	2000	0
Eritrea	2001	0
Eritrea	2002	-1
Changes		27

Table 4-11. Errors in the Logistic, PCA score, Refugees per Capita Model

Note in Table 4-9 that Djibouti's refugees per capita prediction consistently overestimates the expected number of refugees per capita (11, 13, and 11), while Kenya is the source of the most false negatives (-6 and -8). A possible source of these false negatives is the violence in the run up to Kenya's 2002 elections (Adebayo, *et al*, 2002: 17). This also follows the hold out predictions of Kenya for battle deaths, further reinforcing that something about the structure of Kenya has changed recently. Appendix U shows where the errors were occurring, in terms of thresholds. The upper bound of thresholds tested was 7200, yet the model made no errors above 1429. In other cases, such as Kenya in 2006, a false negative is given between the thresholds of 24 and 46 refugees per million, but the model correctly predicts elsewhere. This indicates the underlying structure of the model is changing as the threshold for a positive or a negative result changes, thus the structure of training data was altered as the threshold changed. Djibouti's false positives stem greatly from its falling GDP per capita, especially in relation to the rising GDP of some of the oil rich countries in this study.

The *t*-test scores and their associated *p*-values for the logistic regression model of refugees per capita were explored at several thresholds to help understand how the structure of the model changed as the threshold changed. Based on the results of the Figure 4-8, the *t*-test scores and *p*-values at 0, 450, and 7200 refugees per million were examined. It also brackets the area where the model has the worst performance. The reasoning behind these selections is that the model performs well, both in terms of the training and hold out APER, in each of these regimes.

Referring back to Table 4-2 and 4-8 above, at each threshold examined time effects (PC5) tested as the most significant. That year was such a decisive factor in the model is unsurprising, since refugees per capita throughout the region generally trended upwards during the training data time period (1979-2002). However, the total number of refugees in the region has remained relatively constant, indicating that countries with small populations have been producing disproportionately more refugees than other nations. Thus, the average number of refugees per capita for each country has risen, while the total number of refugees has remained relatively steady. Eritrea, with its war of independence and small population, along with Somalia's disintegration, are the causes of the spike occurring in the late 80's and early 90's.



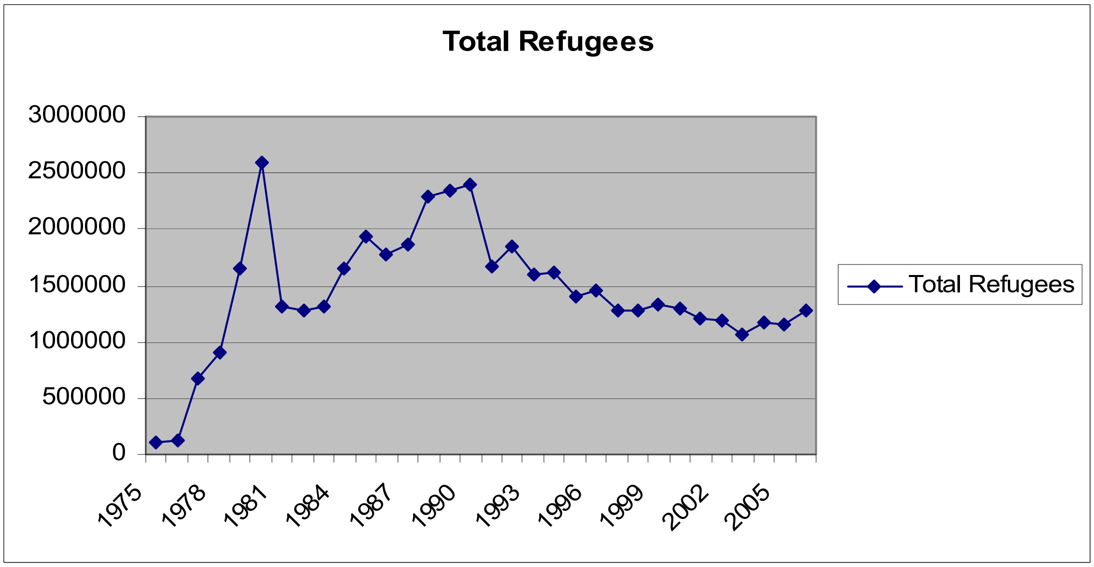


Figure 4-9. Average Refugees Per Capita and Total Refugees in the HoA Countries

Examining the model with a threshold of 450 refugees per million population it can be seen that weak government and existing instability (PC 9) is the next best predictor of refugees four years hence. Note however that weak government and existing instability are not significant to the model at a threshold of zero refugees. This makes some intuitive sense, since the variables loaded on weak government and existing instability represent current instability indicators. If there are zero refugees currently, then there is also more likely to be less battle deaths due to their correlation, and other factors will need to be used to pick out if there will be a refugee problem in 4 years, i.e. when there is a refugee problem now, there is a strong chance there will still be one in 4 years. If there is no refugee problem now, then something else will have to cause a problem 4 years from now. This suggests the situation which started the refugee problem in the region is not being resolved.

In the same vein, agricultural reliance and stress (PC 3) and economic integration into the global market (PC 7) are significant to the 450 and 7200 threshold models, but less so to the model with 0 as a threshold. PC 3 has percentage of paved roads, land stress, arable land per capita, and the water / agriculture as percent of GDP / arable land per capita interaction loaded on it. As these increase, the expected number of refugees does as well. Trade ratio, foreign aid as a % of GNI, and literacy are weighted on PC 7. A positive trade ratio and increased literacy decrease the expected number of refugees per capita, while an increase in foreign aid as a percent of GNI increases the expected number of refugees per capita. The significance of economic integration in the model appears to support Barnett's conclusion's regarding an integrated economy being vital to stability, as well as the negative influence of foreign aid (Barnett, 2003).

Conversely, fractionalization (PC 2), total population, military spending and health statistics (PC 6), and improving quality of life are significant to the model when the threshold is 0, but not at the other two thresholds examined. Gender parity (.3808), ethnic fractionalization (.4409), linguistic fractionalization (.3523), and Kenya (.4008) are the most heavily loaded variables on PC 2. As each of these increase (or become "true" in the case of the binary Kenya variable), the probability that the number of refugees per capita will exceed the threshold increases. This suggests that in the absence of existing war or refugee problems, the variables above become more significant as predictors. Particularly noteworthy, in light of recent events, is that the binary country indicator variable of Kenya is significant when existing refugee levels are low. This suggests that there is some quality about Kenya not reflected in the other data which has heretofore prevented Kenya from experiencing a refugee instability event. Note from Table 4-9 the model has consistently underestimated Kenya in the most recent years in the hold out data (2005 and 2006). One interpretation of this is that whatever was encompassed by the "Kenya" variable before, has changed for the worse.

In conclusion to section 4.3.3, the recommended model for forecasting refugees per capita is using 54 variables to produce 13 rotated principal components and logistic regression to forecast refugees per capita. The model displays a high degree of robustness based on the thresholds examined, displaying an overall hold out data APER of only 0.0559, and its errors are strongly biased towards the preferred false negatives. The model can be implemented by analysts with any access to a statistical software package, and the data is unclassified. Besides providing classification forecasts, the model has also provided additional evidence that the underlying structure defining Kenya's predictions has changed, reinforcing the conclusions from the battle deaths per capita section. The model has also shown that what defines the predictions of refugees per capita changes greatly as the threshold is moved, potentially indicating different causes for different levels of refugee problems. To a member of an embassy staff who is perhaps without support for detailed continual statistical analysis, this study suggests seeing a decreasing trade ratio or a negative change in infant mortality rate in the HoA region may be an indicator that a refugee problem will be going from manageable to unmanageable within the next few years.

4.3.5. Discrete Models of Genocide / Politicide

The same modeling techniques that were considered for battle deaths and refuges were applied to genocides, including: DA, logistic regression, raw data, normalized raw data, PCA scores, normalized PCA scores, raw data as a dependent variable, and per capita data as a dependent variable. The best model both in terms of lowest APER on the hold out data, least false positives, and least variance was using DA on rotated PCA scores and raw genocide deaths data. The APER was .03874, and the variance was 0. The model displayed a bias in errors towards false positives; in 4 out of 5 data sets there were no false positives. This indicates the model always predicted genocides when present, and did not underestimate their severity at any time with 80% of the data sets used. Appendix S shows the model forecast results.

The threshold scale for genocides using raw data was from 0 to 192000, with the threshold incremented by 250 with each iteration. This range is somewhat larger, relatively, than the threshold ranges used for battle deaths and refugees. It can be seen from Figure 4.6 that the threshold of 192000 is clearly to the right of the point of inflection, unlike the previous thresholds used with battle deaths and refugees. It is also far above any politically acceptable level. In fact, the most reasonable threshold based on Figure 4.14 is 0, or near 0 as suggested by the PITF study. However, in the hold out data set only Sudan has instances of genocide. The high maximum threshold was explored to ensure that the model was detecting the genocides and estimating the magnitude of the genocide, rather than simply identifying the country of Sudan based on the input data.

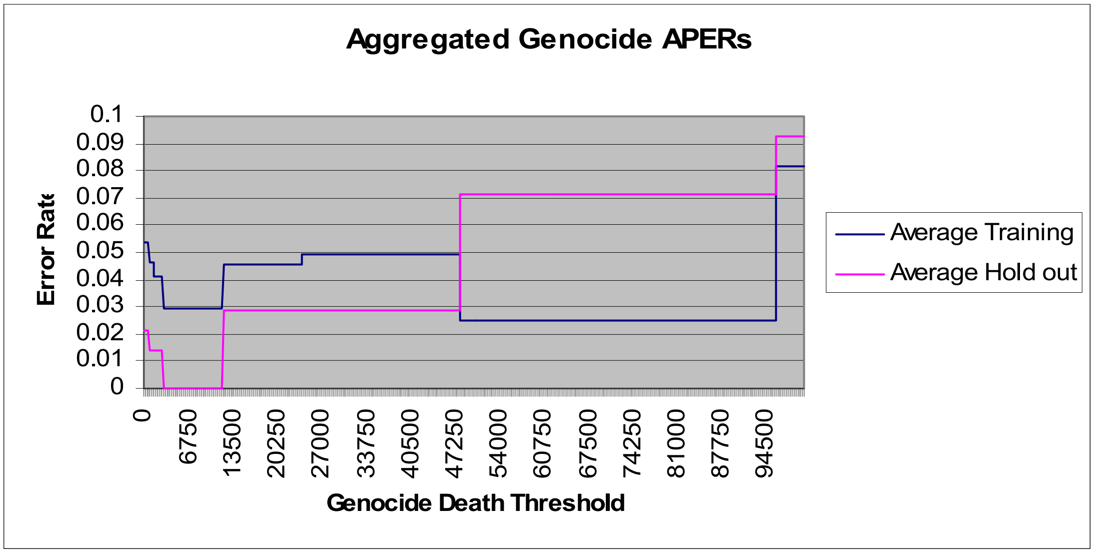


Figure 4-10. Aggregated Genocide APER using DA and PCA scores

The aggregated APERs for the best case model are shown in Figure 4-10. Table 4-12 shows the predictions for each country and year at several thresholds. A 1 indicates that the model predicts a genocide exceeding the threshold at the bottom. The model's prediction errors only change when a new threshold is shown. Thus, from a threshold of 0 to 192000 the model predicted Sudan would have a genocide exceeding the threshold. Once the threshold reached 192000, the model no longer predicted the genocide would exceed the threshold. This produced the false positives in the model, since there were "only" 96000 genocide deaths in Sudan in 2003, and 48000 in 2005 and 2006. This data shows the model is at least detecting the magnitude of genocide, and not just picking Sudan regardless. The error portion of the table indicates where the model predicted higher than the actual number of genocide deaths. Zeros in the error columns indicate correct predictions. One way of interpreting Table 4-10, is that for each exemplar of Sudan, the model has a genocide deaths expected value of somewhere between 96000 and 192000.

		Predictions				Errors			
Yemen	1999	0	0	0	0	0	0	0	0
Yemen	2000	0	0	0	0	0	0	0	0
Yemen	2001	0	0	0	0	0	0	0	0
Yemen	2002	0	0	0	0	0	0	0	0
Somalia	1999	0	0	0	0	0	0	0	0
Somalia	2000	0	0	0	0	0	0	0	0
Somalia	2001	0	0	0	0	0	0	0	0
Somalia	2002	0	0	0	0	0	0	0	0
Sudan	1999	1	1	1	0	0	0	1	0
Sudan	2000	1	1	1	0	0	0	0	0
Sudan	2001	1	1	1	0	0	1	1	0
Sudan	2002	1	1	1	0	0	1	1	0
Djibouti	1999	0	0	0	0	0	0	0	0
Djibouti	2000	0	0	0	0	0	0	0	0
Djibouti	2001	0	0	0	0	0	0	0	0
Djibouti	2002	0	0	0	0	0	0	0	0
Kenya	1999	0	0	0	0	0	0	0	0
Kenya	2000	0	0	0	0	0	0	0	0
Kenya	2001	0	0	0	0	0	0	0	0
Kenya	2002	0	0	0	0	0	0	0	0
Ethiopia	1999	0	0	0	0	0	0	0	0
Ethiopia	2000	0	0	0	0	0	0	0	0
Ethiopia	2001	0	0	0	0	0	0	0	0
Ethiopia	2002	0	0	0	0	0	0	0	0
Eritrea	1999	0	0	0	0	0	0	0	0
Eritrea	2000	0	0	0	0	0	0	0	0
Eritrea	2001	0	0	0	0	0	0	0	0
Eritrea	2002	0	0	0	0	0	0	0	0
Genocide Threshold		0	48000	96000	192000	0	48000	96000	192000

Table 4-12. Genocide Model Results

		Actual	
		Over	Under
Predicted	Over	5845	2175
	Under	0	48120

Table 4-13. Recommended Genocide Deaths Model Confusion Matrix

Note: A 1 in the prediction column indicates the forecast exceeds the threshold at the bottom. A 0 indicates the predicted number of genocide deaths is less than the threshold.

In the errors section, a 1 indicates a false positive (the model predicted more genocide deaths than actually occurred).

Threshold	0	3000	48000	192000
Wealth and Urbanization	0.0275	-0.0199	-0.0154	0.0462
Fractionalization	-0.042	-0.0463	-0.0801	0.1062
Agricultural Reliance / Stress	-0.618	-0.5413	-0.6788	-0.0246
Total Population, Ethiopia	0.0599	0.0434	-0.0484	0.0415
Time Effects	0.2186	0.2156	0.2583	-0.8516
Military / Health Spending	-0.1621	-0.0348	0.0181	-0.0494
Economic Integration into Global Market	0.213	0.0965	0.151	0.0517
Long Term Government Stability	-0.0747	-0.0182	0.0994	-0.1661
Weak Government / Existing Instability	-0.1867	-0.18	-0.1757	0.0706
Change in Infant Mortality Rate	-0.2736	-0.2887	-0.4532	0.2553
Political Discrimination / Exclusion	-0.1776	-0.1791	-0.2008	0.2692
Improving Quality of Life	0.1807	0.1488	0.2818	0.1176
Transition Government	-0.0956	-0.0582	-0.0878	0.1979

Table 4-14. Scaled Discriminant Loadings For Genocide Models Using MI 2

The discriminant loadings shown in Table 4-11 were calculated as per section 3.5.4 Just as in the models for refugees and battle deaths, the principal components most significant to the model change as the threshold used to determine which set the observations belong in changes. Discriminant loadings are interpreted much the same way PCA loadings of standardized data are as scores between -1 and 1 which show how each variable is loaded on the components. In this case, the loadings show how much separation each principal component provides, where a -1 or 1 indicates a principal component perfectly separates the groups in one plane. Given that 192000 genocide deaths is far beyond any tipping point, the following discussion focuses on the discriminant loadings shown above at thresholds of 0 and 3000. Since all of the data was standardized prior to PCA, the relative linear weights of the loadings are indicative of each principal components effect on the model. At thresholds of 0 and 3000 agricultural reliance and stress (PC 3) is the most important variable. The raw variables most heavily loaded on PC 3 are forested land (.3139), percent paved roads (.3572), arable land per capita (.3644), land stress (.3767), and the water per capita/ agriculture as a percent of GDP/ arable land per capita interaction (.3597). As each of these variables increase, so to does the probability of a genocide. This suggests that competition for scarce resources in a subsistence economy is the most significant influence on committing genocide, which supports Collier and Hoeffler's conclusion of economic competition as a primary cause of conflict in "Greed and Grievance"(Collier and Hoeffler, 2002: 16). Given how much stronger PC 3's influence is than any other, it strongly suggests that the key to preventing genocide in the HoA region may be to increase agricultural yields, reduce reliance on agriculture, and increase arable land through irrigation.

The next most significant principal component was change in infant mortality rate (PC 10), which had change in infant mortality rate strongly loaded on it (.7118). Economic integration into the global market (foreign aid as % of GNI and trade ratio) as part of PC 7 had some effect on the model. It is noteworthy how strongly dominated by agricultural factors this model is, which tends to lend credence to Collier and Hoeffler's competition models over the PITF model which places great emphasis on the type of government.

In summary of section 4.3.4 the model recommended for forecasting genocide in the HoA region is using PCA to generate scores which are uses as variables in discriminant analysis. This model has displayed both accuracy (.0387 APER) and a strong bias towards false positive errors. Unlike the previous two models of battle deaths per capita and refugees per capita, a specific threshold of 0 is suggested to follow on researchers and forecasters using this model. The discriminant loadings again suggest that change in infant mortality rate is a single variable that can be used to gauge the potential onset of genocide. However, the variables associated with agriculture are weighted far more heavily, and suggest that an analyst should look at increasing scarcity of resources in an agriculturally reliant society as being a strong predictor of genocide in the future.

4.3.6. Comparing Discrete Models

The most directly comparable model of instability indicators to this study is the PITF model. However, the PITF Phase V documentation states that:

> That is, a relatively simple model (six variables, with no interactions) has remarkable accuracy in distinguishing cases of instability from stable cases. Indeed, the postdictive accuracy of this model is striking: 87 to 92 percent of cases are correctly classified when we choose a cut point that balances model sensitivity and specificity. (Goldstein, *et al*, 2006: 28)

Note that the model above was predicting only whether an instability event would occur, and not what type, with 87-92 percent accuracy. The PITF group also selected "cut points" (thresholds) that were based on their model as well as utility, much the same way a threshold points were selected based on a combination of utility and model capability for this study. Additionally PITF does not predict instability events after a government has already failed, or is in transition (e-mail with Dr. Jay Ulfelder). The PITF predictions also used data that was lagged by 2 years, vice 4 in this study. Accordingly, predictions of stability were generated using the hold out data in the best discrete models with thresholds of .00001 battle deaths per capita, .000450 refugees per capita, and 0 genocide deaths. If a country experienced no instability events as defined by being below the thresholds in each of the instability indicators (battle deaths per capita, refugees per capita, genocide deaths), then it was scored a 0. If there were any instability events, i.e. one of the dependent variables exceeded the pre-set threshold, then the "truth" model was scored a 1. This created a "truth" value with which to compare the predictions of each of the discrete models developed in this study. Then, using the three thresholds described above (.00001, .00045, and 0 respectively) and the models described in sections 4.3.2, 4.3.3, and 4.3.4, hold out data was used to create forecasts for battle deaths per capita, refugees per capita, and genocide. If any of the forecast values for a country predicted an instability event, then that country and year were scored as a 1, just as was done with the actual data. Using the models this way, the models predicted all but Kenya in 2006 correctly, for a correct prediction rate of approximately 96.4%. If the threshold for battle deaths is set any higher than .00001, the error rate drops to 0. Individually, the error rates of models of battle deaths per capita, refugees per capita, and genocide developed in this research have already been described in the previous sections. Only the error rate of battle deaths is greater than 8-13% percent error rate described by PITF. Given these thresholds and models, the results are shown below, along with the "truth" values and the instability prediction classification error.

		Predicted				Actual				
		Battle Deaths	Refugees	Genocide	Instability	Battle Deaths	Refugees	Genocide	Instability	ERROR
Yemen	1999	0	0	0	0	0	0	0	0	0
Yemen	2000	0	0	0	0	0	0	0	0	0
Yemen	2001	0	0	0	0	0	0	0	0	0
Yemen	2002	0	0	0	0	0	0	0	0	0
Somalia	1999	0	1	0	1	0	1	0	1	0
Somalia	2000	0	1	0	1	0	1	0	1	0
Somalia	2001	0	1	0	1	0	1	0	1	0
Somalia	2002	1	1	0	1	0	1	0	1	0
Sudan	1999	1	1	1	1	1	1	1	1	0
Sudan	2000	1	1	1	1	1	1	1	1	0
Sudan	2001	1	1	1	1	1	1	1	1	0
Sudan	2002	1	1	1	1	1	1	1	1	0
Djibouti	1999	0	0	0	0	0	0	0	0	0
Djibouti	2000	0	0	0	0	0	0	0	0	0
Djibouti	2001	0	0	0	0	0	0	0	0	0
Djibouti	2002	0	0	0	0	0	0	0	0	0
Kenya	1999	0	0	0	0	0	0	0	0	0
Kenya	2000	0	0	0	0	0	0	0	0	0
Kenya	2001	0	0	0	0	0	0	0	0	0
Kenya	2002	1	0	0	1	0	0	0	0	1
Ethiopia	1999	1	1	0	1	1	1	0	1	0
Ethiopia	2000	1	1	0	1	0	1	0	1	0
Ethiopia	2001	1	1	0	1	0	1	0	1	0
Ethiopia	2002	0	1	0	1	0	1	0	1	0
Eritrea	1999	1	1	0	1	1	1	0	1	0
Eritrea	2000	0	1	0	1	1	1	0	1	0
Eritrea	2001	0	1	0	1	1	1	0	1	0
Eritrea	2002	0	1	0	1	1	1	0	1	0

Table 4-15. Forecast Errors on Hold Out Data and Stability Classification

Thus, the methods put forward by this study seem to surpass the PITF model in several areas:

1. Increase in lead time by two years.
2. Provides predictions specific to each type of instability event
3. Provides a predictions of general instability at least as accurate as PITF’s

Additionally, each of the individual instability indicator models except battle deaths enjoys a high degree of accuracy, from 92 to 96 percent. Even in the case of battle deaths, the gain of two years would seem to outweigh the small loss of accuracy, which is 83% vice the 87-92 overall accuracy quoted by PITF. The weakness of the methods proposed here are that these techniques require more data, and more data manipulation. Further research may discover ways to simplify these models.

Many of the papers reviewed in the literature review ascribe different causes or indicator variables to instability events. Given the study's best discrete models of battle deaths, refugees, and genocide, the following principal components were found to be influential on each, and are listed in decreasing order of significance from left to right on each model:

Model	Significant Principal Components							Threshold
Battle Deaths per capita	7	10	3	8	1			1.00E-04
Refugees Per Capita	5	9	3	8	1	7	10	0.00045
Genocide Deaths	3	10	5	7	12	11		0

Table 4-16. List of Influential Principal Components on Selected Models

The suggested interpretation of each of the significant principal components listed in Table 4-16 have been previously shown in table 4-2. The threshold column in Table 4-13 shows the threshold used to obtain the model parameters for the three discrete instability forecasting models recommended in this chapter. These thresholds for battle deaths per capita, refugees per capita, and genocide deaths are the same ones used to generate Table 4-12. Note that agricultural reliance and stress (PC 3) is significant to all of the models of the HoA region listed above, suggesting agricultural stressors such as lack of water and arable land is important to forecasting stability in general. Some previous researchers (Goldstone, *et al*, 2005: 21) had actively described agricultural data as inconsequential to their global forecasting models, and claimed that there is no substantive difference when analyzing Africa, but the results of this study suggest otherwise. Economic integration into the global market (PC 7) is also significant to all three models, although not quite as highly ranked across the board as the agricultural factors. This seems to support Barnett's assertion that integration into the global market is key to stability (Barnett, 2003). Change in infant mortality rate, PC 10, though less strongly influential, is significant to all three models. This tends to confirm that previous research that used this statistic as a single warning indicator (O'Brien, 2002: 3).

Other interesting clues to causation exist in the differences between the three models. Political discrimination and improving quality of life (PC 11 and 12) are only significant to the model of genocide, and are the only PCs which are significant to only one model. This implies a combination of political, social, and economic act as conditions which promote genocide. It makes intuitive sense that a minority group which is discriminated against and prevented from participating in government would be in even greater jeopardy if resources were scarce and quality of life were not improving for the government's existing power base.

Another case of differences occurs where a principal component is loaded on only two out of three models. Wealth and Urbanization (PC 1) and Long Term government Stability (PC 8) are only loaded on the models of battle deaths and refugees. Both make sense as reasons to be less prone to going to war and producing refugees, and less likely to suffer from anarchy or transitional governments. Indeed, the signs of the loadings and the associated variables show that as wealth, urbanization, and government longevity increase Conversely, it can be posited that these factors do not have an effect the genocide model because a government which has not had time to consolidate power is less able to enact a genocide, but once it has, it can do so at any time. A potential explanation for PC 1 not being part of the genocide model is that killing unwanted populace can just as easily happen in cities as it can in the countryside. As for wealth, it appears that physical resources such as arable land seem to drive genocide more than purely commercial wealth. Of note is that the variable for principal commodity exports as a percent of GDP was not heavily weighted on any of the principal components.

PC 9, which seems to indicate weak forms of government, current war, and refugees, is significant to the models of genocide and refugees but not of battle deaths. This makes sense because refugee figures have persistence, i.e. if a person is a refugee one year, unless they go home they will be a refugee next year. However, once a person abandons their property it is very difficult to return, as can be seen in Iraq, Israel, and the Palestinian territories (Artz, 1997: 17). Thus refugees tend to carry over for a long time, so the number of refugees today strongly influences the number of refugees tomorrow. Wars, however, tend to produce new deaths over a shorter period, and those deaths do not translate from year to year.

4.4. Predicting 2007 -2010 in the Horn of Africa Region

The purpose of this section is to provide predictions of battle deaths per capita, refugees per capita, genocides, and malnutrition in the Horn of Africa using 2003-2006 input data and the models developed for each in sections 4.2 and 4.3. Malnutrition was a continuous value prediction using OLS regression. Battle deaths per capita and refugees per capita were forecast using rotated PCA scores and logistic regression. Genocide deaths were predicted using rotated PCA scores and DA. The last three were all discrete forecasts. The thresholds used in the discrete models were the same as used in the comparative model analysis in section 4.3.5, and Tables 4-12 and 4-13. The models were run again, using the all the data available since the study has already shown the date structure may change over time. The results of re-running the models are shown in Table 4-14 below. In addition, given that the battle death data and refugee data is on a per capita basis, the threshold exceeded in raw form is shown.

Country	Year	Battle Deaths Per Capita	Refugees Per Capita	Genocide	Malnutrition	Instability
Yemen	2007	0	0	0	22.5317	0
Yemen	2008	0	0	0	19.1	0
Yemen	2009	0	0	0	14.4285	0
Yemen	2010	0	0	0	9.1343	0
Somalia	2007	0	1	0	18.2328	1
Somalia	2008	0	1	0	9.7749	1
Somalia	2009	0	1	0	5.9229	1
Somalia	2010	0	1	0	19.1541	1
Sudan	2007	1	1	1	15.6679	1
Sudan	2008	1	1	1	10.6253	1
Sudan	2009	1	1	1	7.1761	1
Sudan	2010	1	1	1	3.8282	1
Djibouti	2007	0	0	0	29.891	0
Djibouti	2008	0	0	0	26.824	0
Djibouti	2009	0	0	0	17.0651	0
Djibouti	2010	0	0	0	17.2837	0
Kenya	2007	0	0	0	24.7594	0
Kenya	2008	0	0	0	18.8233	0
Kenya	2009	0	0	0	10.6466	0
Kenya	2010	0	0	0	5.9641	0
Ethiopia	2007	0	1	0	37.724	1
Ethiopia	2008	0	1	0	38.8035	1
Ethiopia	2009	0	1	0	32.5342	1
Ethiopia	2010	0	1	0	29.6439	1
Eritrea	2007	1	1	0	65.9558	1
Eritrea	2008	1	1	0	66.7145	1
Eritrea	2009	1	1	0	58.8768	1
Eritrea	2010	0	1	0	67.2961	1

Table 4-17. Forecasting Intability in the HoA Region 2007-2010

The data was also examined to determine what range of raw data the positives fell in for battle deaths and refugees.

When the change logs were examined, the range in which positives were occurring was also observed. In the case of Sudan, the model showed Sudan will continue to experience high levels of violence, but those levels will gradually decline going into 2010, with a the maximum threshold for a positive declining from 22577 total battle deaths in 2007 to 11919 in 2010. The increasing oil wealth of Sudan and increasing standard of living tends to reduce the predictions for Sudan's battle deaths, despite its violent history. Eritrea is expected to have low level violence, with total deaths remaining below 60 battle deaths per year, and declining as time goes by, to the point where no battle deaths are expected in 2010. Ethiopia is the most difficult to predict. Ethiopia comes very close to exceeding the threshold of .00001 (it gives a positive result all the way up to .0000085). Based on the structure of this run, Ethiopia seems most likely to have provided a false negative, whereas Eritrea seems most likely to have given a false positive if the threshold is set at .00001. The relative calm between Ethiopia and Eritrea, as well as their governments building up "durability", tends to improve their forecasting outlook. Also, trends away from agricultural reliance aid their predictions.

The predictions for refugees provides few surprises. The countries that have refugee problems today, Sudan, Somalia, Ethiopia, and Eritrea in particular, are expected to continue to have refugee issues, and the model projects they will get worse in some cases before they get better. When examining the predictions along the range of thresholds, not only do Sudan, Somalia, and Eritrea exceed the nominal threshold of .00045 per capita, they exceed it all the way out to .007. Ethiopia exceeds the nominal threshold as well, but when the actual predictions are observed along the spectrum of thresholds, Ethiopia's refugee problem reaches .0007 per capita, and is increasing from 2007 to 2008. Yemen, however, provided some positive results over a number of spectrums, although not in the nominal one (.00045). This indicates Yemen is displaying some, but not all of the traits associated with producing refugees. As the PC loadings on the model change along with threshold, however, Yemen's prediction changes somewhat irregularly. This may in part be due to Yemen being structurally different from the other countries in the region in terms of culture, geography, and social homogeneity. Without some sort of intervention or structural change, the refugee crisis in the HoA region is only expected to intensify in 2008 and going forward.

The predictions for genocide continue to forecast the government of Sudan will continue to engage in a policy of altering demographics of Darfur and the oil rich Southern Sudan to suit their aims. No positives were given for other countries, and the magnitude of the expected genocide in Sudan did not decrease based on the changes in the threshold. This is expected to be an issue future administrations will have to deal with as well.

Undernourishment was predicted to fall across the region, with the exception of Eritrea. The latter is particularly disturbing, given the correlation between undernourishment, agricultural factors, infant mortality, and the other indicators of instability. Of the problems facing the HoA region, this is one of the few that has some potential to be avoided by early intervention. There are no current hostilities in Eritrea, and it has access to deep water ports.

4.5 Analysis Conclusions and Summary

Numerous continuous models were attempted, however the only one which seemed to sufficiently explain variance in the model, while having a small enough RMSE to be useful as a predictive tool was using OLS regression to forecast undernourishment. Discrete models were developed to predict battle deaths per capita, refugees per capita, and genocide deaths. For the first two, the preferred model was 13 PCA scores generated from 54 variables discriminating with logistic regression to develop over / under predictions against a range of thresholds. For genocides similarly developed PCA scores were instead used on raw genocide data, and the preferred method of classification groups was found to be discriminant analysis. The results of these models, and their overall ability to predict instability events with a 4 year lead time was at least equal to previous models, if not exceeding their accuracy in terms of forecasting refugees and genocide deaths.

This study recommends the use of the four models described in this chapter. The general structure of the models for undernourishment, battle deaths per capita, refugees per capita, and genocide deaths can be summarized as:

1. Continuous forecasting model of undernourishment for each country in the Horn of Africa region using canonical correlation scores and Ordinary Least Squares regression.
2. Discrete forecasting model of battle deaths per capita for each country in the Horn of Africa region using principal component scores and logistic regression.
3. Discrete forecasting model of refugees per capita for each country in the Horn of Africa region using principal component scores and logistic regression.

4. Discrete forecasting model of genocide and politicide for each country in the Horn of Africa region using principal component scores and Discriminant Analysis (DA).

Model	Data Transformation	Continuous / Discrete	Model Type	Section	Page
Undernourishment	Canonical Correlation	Continuous	OLS Regression	4.3.1	4-5
Battle Deaths Per Capita	Principal Components	Discrete	Logit Regression	4.3.3	4-16
Refugees Per Capita	Principal Components	Discrete	Logit Regression	4.3.4	4-27
Genocide / Politicide Deaths	Principal Components	Discrete	Discriminant Analysis	4.3.5	4-37

Table 4-18. Summary of Model Types and Locations of Descriptions

Table 4-18. lists the each model, how the data was transformed to make it better structured for the mathematical techniques used, whether the resulting model was continuous or discrete, and what mathematical technique was used to model the dependent data. Additionally, the location and section where each of these models is presented is listed on the right.

These models are recommended based on their accuracy, robustness, flexibility, error biased against false negatives, and ease of application of each of these models by analysts. Each model provided insight into the potential causes of instability indicators in the HoA region, as well as suggesting the trackable indicator variables of change in infant mortality rate and trade ratio.

When reviewing the loadings of variables on PCA scores and how the PCA scores were loaded on each model, it was found that no one model proposed by subject matter experts proved to be dominant; and that in fact the data supported more than one theory, and that no one theory explained everything within the results. Using the model to predict instability events in the 7 countries in this study from 2007-2010 did not show any great surprises. The countries wracked with war, refugees, and genocide continued to be unstable, although the model predicts the overall level of battle deaths will taper off by about half by 2010 due to reduced agricultural reliance and greater international trade in the region. Refugees will be the dominant problem, particularly from Eritrea, Somalia, and Sudan, and will still be there in 2010. Genocide will continue to be limited to Sudan, according to the model. The models of Kenya for 2003-2006 were consistently underpredicting instability indicators, suggesting that the country's most recent troubles are part of a continuing destabilization.

Chapter 4 has developed, presented, discussed, and recommended 4 models of instability indicators in HoA region. Chapter 5 will discuss the research contributions of this study and areas of future research suggested by this work.

5 Conclusions and Recommendations

5.1 Introduction

This Chapter concludes this study with a summary of significant findings and contributions to the field as well as recommendations for future research.

5.2 Research Contributions

This thesis provided description of the development and use of mathematical models for forecasting the instability indicators of battle deaths per capita, refugees per capita, genocide and politicide deaths, and undernourishment in the Horn of Africa region. It extends the prediction interval from two years to four, while maintaining a forecast error rate for each of the indicators and the overall predictions of instability events on a par with, or even better than current models. The results support numerous theories regarding the causes and leading indicator variables associated with instability events, while at the same time refuting that any one studies' group of variables are the only ones necessary for building a model. The research supports further multivariate study, rather than looking for a simplified "one size fits all" model. The methods used in this study demonstrate the feasibility of longer term predictive models of specific types of instability, which will allow policy makers more time to act accordingly.

5.3. Recommendations for Action

Those interested in this research should strongly consider using the forecasting models of undernourishment, battle deaths per capita, refugees per capita, and genocide recommended in sections 4.3.1, 4.3.3, 4.3.4, and 4.3.5 of this study to forecast stability conditions in countries in the Horn of Africa region. The discrete models provide predictive accuracy between 83 and 96%, as well as a longer 4 year lead time than current models. The continuous model of undernourishment provides the same lead time, and displayed $R^2_{\Pr ed}$ of .9217, an R^2 of ..9152, and an RMSE of 4.9. The recommended models are implementable with commercial off the shelf software, and the data is readily available. Missing data can also be dealt with relatively easily using vetted freeware. However, when it is not possible to utilize the full models presented here, certain variables can serve as general early warning indicators for the countries in the HoA region. Agricultural statistics, infant mortality rate, and trade ratio are dominantly loaded on the three principal components which are all significant to each of the three discrete models and have a high rate of availability.

Given the forecasts in section 4.4 and shown in Table 4-14, there are certain countries which have the potential to descend into general instability. Djibouti, the site of a U.S. military installation, displays some of the warning signs of instability, particularly the decreases in GDP per capita and its reliance on foreign aid put it at risk for a refugee problem and to a lesser extent armed conflict. Eritrea's projected increases in undernourishment while all the other countries are forecast to gain in that regard is troubling. Famines are typically accompanied by a refugee crisis, and the region has seen this phenomenon before in 1984-1985 (BBC News, 2000). Given the strong French and U.S. military presence in Djibouti, and the lack of current violence, there may be an opportunity to take steps to modernized Djibouti's infrastructure, particularly that of transportation. Namsuk Cho's study suggests that before anything else can be done to re-build, or build up, an urban area like Djibouti, the roads and other transportation infrastructure items must first be re-built. Given a stronger infrastructure, and it's location at the mouth of the Red Sea, Djibouti could conceivably position itself as a significant commercial waypoint between Asia and Europe.

Kenya's recent troubles at the beginning of 2008 are the most worrisome. The trend of the model to underpredict the severity of Kenya's battle deaths per capita, refugees per capita, and undernourishment along with the significance of the "Kenya" indicator variable in a principal component associated increasing refugees per capita is extremely troubling. This indicates at some level Kenya has changed, and it is not behaving as it has in the past. Unfortunately, these apparent changes all point to increasing instability. While it cannot be told if the point of no return has been reached, based on historical data it appears Kenya is already in the transition phase. Based on the stability index shown in Chapter 3 when other countries have reached the transition stage, it can be expected 2009 and 2010 could be very bad for Kenya. Since fighting has already broken out, the international community has fewer options when considering how to arrest Kenya's slide. A negotiated political settlement, even a temporary one, between the two primary factions in Kenya might buy some cool down time as well as an opportunity for fair, internationally observed elections that will settle the power struggle in a way that both parties will have to accept if they really do want democracy. There are also ways to economically incentivize democracy for both groups in Kenya.

In general, assuming that there is some causality associated with the principal components involved in these models, agriculture is the key to stabilizing the Horn of Africa. In order to attack the problem, a two pronged approach is needed. The first is to reduce a country's reliance on agriculture via the fostering of alternative industries. The other is to reduce the scarcity of agricultural resources, particularly irrigated arable land.

It is beyond the scope of this study to determine how these effects should be accomplished, however the data has strongly suggested undernourishment, refugees, war, genocide, and agricultural factors are all linked to each other.

5.4. Recommendations for Future Research

The models in this study are by no means perfect. Further attempts at building a continuous model, as well as building a better discrete model seem to have a strong chance at success, given the wide variety of untried methods found during the course of this study. Additionally, there were tantalizing suggestions that some data hidden within the binary country variables might provide insight and a better, more accurate model. Thus, models with superior accuracy, simpler implementation, longer lead time, more utility to agencies interested in the forecasts, and better interpretability may all be possible. This section briefly discusses some avenues of research which may provide further gains in the field.

5.4.1. Generalized Auto-Regressive Conditional Heteroscedasticity

Generalized Auto-Regressive Conditional Heteroscedasticity (GARCH) models are useful in cases where a model needs to predict a variable that has long periods of relative stability interrupted by shorter highly volatile ones. This technique has previously been used by econometricians to predict volatile markets, but it's ability to forecast variance, suitability for heteroscedastistic problems, and ability to autoregress multiple variables besides the one being forecast to make its predictions suggests GARCH is a potentially useful method for forecasting even more accurately, particularly on continuous variables such as battle deaths per capita or refugees per capita (Enders, 2004: 118). Unlike the models used in this study, this is more truly a time series analysis, vice one where time is simply a variable considered in the data set. Section 4.3.5 provides some clues as to which 4 year variables may be significant to a GARCH model of instability indicators. Change in infant mortality rate, trade ratio, and agricultural stressors and reliance were significant to each of the four models and might be a starting point for GARCH analysis.

5.4.2. Multivariate Adaptive Regression Splines (MARS)

MARS is a highly complex, computationally intensive, non-parametric method of building continuous models using lagged time series data proposed in 1991 by Dr. James Stevens in a dissertation from the Naval Postgraduate School. It is described in his abstract:

> MARS can be conceptualized as a generalization of recursive partitioning that use spline fitting in lieu of other simple fitting functions. MARS is a computationally intensive methodology that fits a nonparametric regression model in the form of an expansion in product spline basis functions of predictor variables chosen during a forward and backward recursive partitioning strategy. The MARS algorithm produces continuous nonlinear regression models for high-dimensional data using a combination of predictor variable interactions and partitions of the predictor variable space. By letting the predictor variables in the MARS algorithm be lagged values of a time series system, one obtains a univariate (ASTAR) or semi-multivariate (SMASTAR) adaptive spline threshold autoregressive model for nonlinear autoregressive threshold modeling and analysis of time series, thereby extending the threshold autoregression (TAR) time series methodology developed by Tong (Stevens, 1991: 1).

This technique represents another potential way of developing a continuous time series model of stability indicators using lagged time series data. The drawbacks are the long run time to test any model, and the relatively low number of observed data points in this study’s training model. However, given more points, and more time to conduct further studies, this technique appears well suited to building a more accurate predictive model of instability indicators.

5.4.3. Three Group Discriminations

Based on the graphs in Figures 4.12 and 4.13, it appears that there may actually be three general groups that countries fall into based battle deaths per capita and refugees per capita, stable, unstable, and failed. The unstable transition zone tends to be the narrowest of the three, but the data in suggest some of the forecasts for 2010 would fall into this critical range. Additionally, the study by the Harmony Project suggests that this is the zone that is most likely to breed terrorism (Harmony Project, 2007: 47). Thus, investigating stability indicators using a 3 group division for battle deaths, refugees, and genocide would provide an increased level of granularity for forecasters to tell which countries are "right on the edge". Given the success of using PCA scores with logistic regression and DA, performing three group classification using the methods demonstrated in this study appears highly likely to provide additional insight and predictive ability. Logistic regression and DA using quadratic discriminant scores are both capable of creating classification models with more than two groupings, as well as predicting into which category new observation will fall.

5.4.3.1 Cluster Analysis (CA)

In addition to using DA and logistic regression to build a three group model, cluster analysis is another method of finding discrete multi-category groupings (Lattin, *et al*, 2003: 264-265). CA creates discrete non-spatial representations of multivariate data. CA tries to find naturally occurring groupings with regions of high observation density surrounded by lower density. This serves to find the clearest possible distinction between groups, and allows new observations to be assigned to the one with the highest probability density in the n-dimensional space. Additionally, the user can pre-define how many groups they wish for the observations to be separated into. The process of forming clusters is iterative, where within group variation is compared with between group variation and new members are reassigned until the between group variance is minimized (Dillon, 1984: 160). The drawback to CA is that the user does not define what category each cluster represents, and thus the contents of the clusters are subject to interpretation.

5.4.4 Modifying the Data Set

This study has shown there is a benefit to using data from the entire region to create models which are then used to create individual country forecasts of stability indicators. The model results shown in Appendix R used only data from their individual countries. These models in Appendix R fared poorly when compared with models using data for all seven countries used in this study, however. The study has also shown some benefits when using more observations to reduce variance, particularly in the continuous models. Additionally, the significance of country indicator variables suggests there are other variables which these are proxies for, and may provide insight into the actual workings behind these proxies. The following sections are suggestions for ways in which the data sets used here could be modified in order to improve both the model's accuracy, as well as it's interpretability.

5.4.4.1. Add More Years of Data

If some of the time series analysis methods described in this chapter are employed, adding more years worth of data to the set would assist in creating a model with more accurately modeled variance, particularly in the case of GARCH which attempts to model the changes in variance over time. Additionally, it would allow the addition of other variables, or testing for interactions between variables, without degrading variance nearly as much.

5.4.4.2. Add and Remove Variables

Some of the variables used in this study did not end up being heavily loaded on any of the principal components retained. Time constraints, and the scope of this study precluded studies prevented investigations of models built by removing the variables not highly loaded on any of the 13 rotated principal component scores. Multicollinearity precluded the use of the religious fractionalization scores in the data. It would be interesting for future research to find a way to remove the other data which caused the mulitcollinearity, and re-insert the religious fractionalization data. Conversely, this study was unable to acquire data on many variables that had been suggested as significant in previous research, such as crime data, the price of an AK-47, GINI coefficients, and many others. The PITF database contains far more variables and years of observations than are contained in this study. Future access to it would only strengthen later modeling attempts.

In some cases, much of the data had to be imputed. One such case was primary commodity exports as a percent of GDP (PCE). Since so much of the data had to be imputed, it is not possible to say with any certainty that PCE is not significant, even though it did not load heavily on and of the principal component scores. Another area for improvement would be if actual data on the number of starvation deaths per year per country could be found and used instead of the less complicated undernourishment data, which was used as a proxy.

5.4.4.3. Add More Countries

The al Qaeda bombing of the U.S. embassy in Tanzania in 1998 demonstrates that other countries in the same general region play important roles in the Global War on Terror. Additionally, the historic roles Chad has played in the Sudanese conflicts and genocides, and the genocides in Uganda and Rwanda all suggest an expanded model would make further predictions more robust. Future studies using the methods and data contained in this research are highly encouraged to consider adding Chad, Uganda, Tanzania, and Rwanda to the data set. Other states of interest worth considering are the Democratic Republic of the Congo and the Central African Republic. Adding these countries could potentially allow for the addition of religious fractionalization to the data set, provide more degrees of freedom for analysis, and further normalize the distribution of the PCA scores data.

5.4.5. Change Point Analysis

There is a sub-set of statistical process control theory which deals with change points. The data in this study suggests that Kenya may have begun to show changes in the structure of its data which reflected destabilization, even though the model continued to under predict the actual outcome. Detecting the change as soon as possible, and being able to differentiate it from simple noise in the data will allow a better understanding of what events constitute sparks for long term problems, as described by the "oily rag" theory. Future research should look into finding change points in large time series multi-variate datasets to identify turning points, and to look for change points that go back further than the 3 to 6 months hypothesized by Gurr and Harff (Gurr and Harff, 1998, 558).

5.4.6. Neural Networks

The PITF Phase V report states "Despite our best efforts….neural networks better predictions of instability…" (Goldstone, *et al*, 2005: 9). However, pilot studies using Generalized Regression Neural Networks (GRNNs) on multiply imputed data set 2 showed an apparent error rate on battle deaths per capita less than that of the logistic regression model recommended in this study. There are many different types of neural networks (radial basis function, feed forward, probabilistic, GRNNs) offering a variety of advantages. Future research into neural nets using the PCA and canonical correlation scores may provide more accurate models for each of the instability indicators.

5.5. Conclusions

The Horn of Africa will continue to be an unstable region in the foreseeable future, and will continue to be subject to the pressures of the international press. Additionally, whether one subscribes to the theory that failed states are havens for terrorist organizations, or near failing states are, the Horn of Africa region will supply both. The inescapable conclusion is that U.S. military and diplomatic intervention in the region is very likely to be necessary in the coming years. It is vital for both to have an accurate model which provides insight into different instability indicators to "lean forward" in anticipation of the predicted instability. The four models presented in this study forecast the types of instability that are likely, they can also be used to estimate the severity. They also provide ample lead time to plan and pre-stage whatever equipment or personnel are deemed necessary. The adage "an ounce of prevention is worth a pound of cure" is appropriate to this situation, and this study provides a new means by which to identify the need for prevention, why the prevention is needed, and insight into what form the prevention should take.

This study showed that the use of multivariate transformation techniques allows for the more effective use of both DA and logistic regression to forecast the instability indicators of battle deaths per capita, refugees per capita, and genocide deaths. In the case of malnutrition, simple linear models were adequate to a 4 year prediction. The discrete forecasts showed strong predictive abilities, and the PCA loadings of the significant variables in each model suggested the underlying structure of each type of instability. Further research into both discrete and continuous models of each of the stability indicators appears to offer a great deal of promise. It is the sincere hope of this study that the models resented will save time, money, and most importantly lives.

Appendix A: Initial Variables In Data Set

This appendix shows the data that was collected for the initial data set. Derived values, such as kilometers of roads per roads per capita, are discussed in Chapter 3. The source refers to where the actual data came from. SME (Subject Matter Expert) code indicates who suggested this variable as a source or indicator of instability. For more detailed descriptions of the sources, see the bibliography.

Source Codes:

1. Food and Agriculture Organization of the United Nations
2. US Census Bureau International Database
3. Freedomhouse.org
4. United Nations Common Database
5.. POLITY IV Data Set
6. Penn World Table
7. World Bank World Development Indicators Database
8. Gathered by Dr. Romain Wacziarg at Stanford University from CIA, Encyclopaedia Brtiannica, Scarrit and Mozaffar, Levinson,World Directory of Minorities, and national census data
9. CIA World Fact Book (1975-2006)
10. USAid DOLPHIN database
11. Center for International Development and Conflict Management
12. Uppsala Conflict Database
13. Minorities at Risk Project
14. Dr. Anke Hoeffler via the World Bank
15. United States Center for Disease Control
16. Center for Global Policy at George Mason University
17. United Nations High Commissioner on Refugees Database

Subject Matter Expert Source:

1. Center for Army Analysis ACTOR Model
2. Political Instability Task Force PITF IV report
3. "Greed and Grievance in Civil War", Collier and Hoeffler's
4. "Classifying Failing States" by Capt. Nysether
5. "Investigating the Complexities of Nationbuilding: A Sub-National Regional Perspective", by Capt. Robbins
6. "Prioritization of Critical Infrastructure Rebuilding" By Capt. Namsuk Cho (ROK)
7. "Nationalism in the Horn of Africa" by Jaquin-Berdal
8. "The Pentagon's New Map" by Barnett
9. Durch Briefing to CJCS
10. "Economic Causes of Civil War" by Collier and Hoeffler
11. "Darfur: A Short History of a Long War", de Waal and Flint

Data Type	Source	SME	Missing
Calories Per Capita Per Day	1	1	20.9%
Infant Mortality Rate	2	1, 2	6.3%
Life Expectancy	2	1	3.4%
Youth Bulge	2	1	20.9%
Gross Domestic Product in Constant 1998 USD	4	1	19.4%
Trade Openness	6	1	16.5%
Total Population	2	7	0.0%
Percent of Population Living in Urban Areas	7	7	0.0%
Telephone and Cell Phone Subscribers per 100 population	4	7	12.6%
Ethnic Fractionalization	8	1	0.0%
Linguistic Fractionalization	8	7	0.0%
Religious Fractionalization	8	1	0.0%
Adult Literacy Rate	4,9,10	7	79.6%
Gender Parity in Elementary Education	10	7	78.6%
Education Spending as a Percentage of Gross National Income	7	3	46.1%
Foreign Aid as a Percentage of Gross National Income	7	8	26.7%
Military Expenditures as a Percentage of Gross National Income	7	7	54.9%
Agricultural Revenue as a Percentage of Gross National Income	7	9	30.6%
Political Discrimination	13	2	24.3%
Economic Discrimination	13	2	24.3%
Executive Recruitment (EXREC)	5	2	0.0%
Government Restrictions on Political Competition (PARCOMP)	5	2	0.0%
Number of Neighboring Countries at War	12	2, 9	0.0%
Years since Last Change in Government	5	8	0.0%
Value of Imported Goods	7	1,8	16.5%
Value of Exported Goods	7	1,8	16.5%
Foreign Aid Per Capita	7	8	4.4%
Primary Commodity Exported as a Percentage of Gross National Income	14	3, 8	86.4%
Annual Gross Domestic Product Growth	7	1,2,3,8,9	26.7%
Percentage of Land that is Forested	7	10	63.6%
Arable Land	7	11	9.7%
1000's of Cubic Meters of Renewable Freshwater Available	4	11	0.0%
Percentage of Roads Paved	7	6	64.6%
Kilometers of Roads	7	6	64.6%
Battle Deaths	12	2	0.0%
Genocide and Politicide Deaths	16	2	0.0%
Net Refugees	17	2	0.0%
Years Since Last Conflict	12	3	0.0%
Percent Undernourished	1	2	81.1%

Appendix B: Derived Variables

Appendix B shows which variables were derived by manipulating the original data via some means, and the nature of those mathematical operations. This does not include transformations, which will be described in Appendix F.

Data Type	Formula
Arable Land Per Capita	Arable Land / Population
Population Density	Total Population / Land Area in Km^2
Water Per Capita	Total Renewable Freshwater Resources in m^2 / Total Population
Trade Ratio	Exports / Imports
Water Per Capita and Agriculture Interaction	Water Per Capita * (100 - Agriculture as a Pct of GDP)
Land Stress	Arable Land Per Capita * (100 - Agriculture as Pct of GDP)
Water, Agriculture, and Land interaction	Water Per Capita * Arable Land Per Capita * (100 - Agriculture as Pct of GDP)
Roads Per Capita	Total Km of Roads / Total Population
Battle Deaths a Pct of Population	(Battle Deaths + Once Sided Conflict Deaths) / Total Population
Genocide and Poiticide Deaths as Pct of Population	(Genocide and Politicide Deaths) / Total Population
Water, Agriculture, and Land interaction	Water Per Capita * Arable Land Per Capita * (100 - Agriculture as Pct of GDP)

Appendix C: Dataset Definitions

The definitions of each variable used, or not previously defined in the text, are listed here in alphabetical order. Unless otherwise noted, the source of the definition is the same as the source of the data. A description of how government types, factionalism, political discrimination, and economic discrimination are defined and scored is provided following the basic variable definitions below.

Variable	Definition
Adult Literacy Rate	The percentage of people ages 15 and over who can, with understanding, read and write a short, simple statement about their everyday life.
Agriculture as Pct of GNI	Agriculture corresponds to ISIC divisions 1-5 and includes forestry, hunting, and fishing, as well as cultivation of crops and livestock production. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator.
Aid as a Pct of GNI	Aid includes both official development assistance (ODA) and official aid. Ratios are computed using values in U.S. dollars converted at official exchange rates.
Aid per capita	Aid per capita includes both official development assistance (ODA) and official aid, and is calculated by dividing total aid by the midyear population estimate.
Arable Land (in Hectares)	Arable land (in hectares) includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded.
Bad Neighbors	The number of countries on a nations border at war in a given year, as recorded by the Upsala Conflict Database

Variable	Definition
Battle Deaths	Counted as battle-related is conflict behaviour between warring parties in the conflict dyad, which is directly related to the incompatibility, , i.e. carried out with the purpose of realizing the goal of the incompatibility and results in deaths. Typically, battle-related deaths occur in what can be described as “normal” warfare involving the armed forces of the warring parties. This includes traditional battlefield fighting, guerrilla activities (e.g. hit-and-run attacks / ambushes) and all kinds of bombardments of military units, cities and villages etc. The targets are usually the military itself and its installations or state institutions and state representatives, but there is often substantial collateral damage in the form of civilians being killed in crossfire, in indiscriminate bombings etc. All deaths – military as well as civilians – incurred in such situations, are counted as battle-related deaths.
Caloric Intake	Estimate of the average number of calories consumed per person per day
Economic Discrimination	Macro codings of the role of public policy and social practice in maintaining or redressing economic inequalities. There are no codes for specific types of restrictions on economic activities. (Specific codes described in main body)
Education as Pct of GNI	Education expenditure refers to the current operating expenditures in education, including wages and salaries and excluding capital investments in buildings and equipment.
Executive Recruitment	Concept variable combines information presented in three component variables: XRREG, XRCOMP, and XROPEN. XRREG is the extent of institutionalization – or regulation – of executive transfers. XRCOMP is the competitiveness of executive selection. XROPEN is the openness of executive recruitment.
Exports of Goods and Services	Exports of goods and services comprise all transactions between residents of a country and the rest of the world involving a change of ownership from residents to nonresidents of general merchandise, goods sent for processing and repairs, nonmonetary gold, and services. Data are in current U.S. dollars.
Forest Area	Forest area is land under natural or planted stands of trees, whether productive or not.
Gender Parity in Education	The ratio of the female-to-male values (or male to female, in certain cases) of net primary school enrollment rates (NER). NER measures the number of pupils in the official age group for a given level of ecudation, expressed as a percentage of the population in that age group.

Variable	Definition
GDP Annual Growth	Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2000 U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.
Government Resrtictions on Political Competition (PARCOMP)	The competitiveness of participation refers to the extent to which alternative preferences for policy and leadership can be pursued in the political arena. Political competition implies a significant degree of civil interaction, so polities which are coded Unregulated (1) on Regulation of Participation (PARREG, variable 2.5) are not coded for competitiveness. Polities in transition between Unregulated and any of the regulated forms on variable 2.5 also are not coded on variable 2.6. Competitiveness is coded on a fivecategory scale.
Gross Domestic Product Per Capita	Annual GDP per person in constant 1998 U.S. dollars.
Imports of Goods and Services	Imports of goods and services comprise all transactions between residents of a country and the rest of the world involving a change of ownership from nonresidents to residents of general merchandise, goods sent for processing and repairs, nonmonetary gold, and services. Data are in current U.S. dollars.
Infant Mortality Rate	Number of deaths of children under the age of 1 per 1000 live births.
Life Expectancy:	Average life expectancy of both males and females at birth.
Military Expenditure as % GDP	Military expenditures data from SIPRI are derived from the NATO definition, which includes all current and capital expenditures on the armed forces, including peacekeeping forces; defense ministries and other government agencies engaged in defense projects; paramilitary forces, if these are judged to be trained and equipped for military operations; and military space activities. Such expenditures include military and civil personnel, including retirement pensions of military personnel and social services for personnel; operation and maintenance; procurement; military research and development; and military aid (in the military expenditures of the donor country). Excluded are civil defense and current expenditures for previous military activities, such as for veterans' benefits, demobilization, conversion, and destruction of weapons. This definition cannot be applied for all countries, however, since that would require much more detailed information than is available about what is included in military budgets and off-budget military expenditure items.

Variable	Definition
Paved Roads (Pct)	Paved roads are those surfaced with crushed stone (macadam) and hydrocarbon binder or bituminized agents, with concrete, or with cobblestones, as a percentage of all the country's roads, measured in length.
Political Discrimination	Macro codings of the role of public policy and social practice in maintaining or redressing political inequalities. (Specific Codes defined in main body)
Primary Commodity Ratio	The total value of exports from a single commodity (coffee, oil, tea, bauxite, etc…) divided by gross domestic product in same dollar terms.
Roads, Total Network	Total road network includes motorways, highways, and main or national roads, secondary or regional roads, and all other roads in a country.
Telephone and Cell Phone Subscibers per 100 population	The number of land line and cell phone telephone accounts per 100 population.
Total Population	The population estimate for July 1 of the given year
Total Renewable Freshwater Respources	Internal flow + Actual external inflow of surface and groundwaters. Measured in millions of cubic meters.
Trade Openness	Value of a country's total imports and exports as a percentage of GDP per capita.
Undernourishment	Undernourishment refers to the condition of people whose dietary energy consumption is continuously below a minimum dietary energy requirement for maintaining a healthy life and carrying out a light physical activity.
Urban Population as Pct of Total	Urban population is the midyear population of areas defined as urban in each country and reported to the United Nations.
Years Since Last Conflict	Number of years since last conflict year as recorded by the Upsala Conflict Database
Youth Bulge:	Ratio of population aged 15-29 to those 30-54.

The governmental categories are defined by:

* Full Autocracy: EXREC<6 & (PARCOMP<3, not equal to 0)
* Partial Autocracy: EXREC<6 or (PARCOMP<3, not equal to 0)
* Partial Democracy with Factionalism: EXREC>=6 & PARCOMP=3
* Partial Democracy without Factionalism: EXREC>=6 & (PARCOMP>3 or =0) & not a full democracy
* Full Democracy: EXREC=8 & PARCOMP=5
* Transitional: Polity = -88 (Dr. Ulfelder correspondence, 10/10/07)

Thus, six independent variables were used to describe the government of each nation in each year. Only one variable could be designated at the high setting of one for each country in each year. Thus, a country with a POLITY IV score of -77 in 1995 would have a 1 in the column for transition government, and 0's in the columns for Anarchy, Full Autocracy, Partial Autocracy, Partial Democracy with Factionalism, Partial Democracy with Factionalism, and Transition Government.

The other categorical variable was the MAR economic discrimination and political discrimination scores. The basic definition of each is in shown below.

POLDIS
0 = No discrimination
1 = Neglect/Remedial policies
Substantial under representation in political office and/or participation due to historical neglect or restrictions. Explicit public policies are designed to protect or improve the group's political status.
2 = Neglect/No remedial policies
Substantial under representation due to historical neglect or restrictions. No social practice of deliberate exclusion. No formal exclusion. No evidence of protective or remedial public policies.
3 = Social exclusion/Neutral policy
Substantial under representation due to prevailing social practice by dominant groups. Formal public policies toward the group are neutral or, if positive, inadequate to offset discriminatory policies.
4 = Exclusion/Repressive policy
Public policies substantially restrict the group's political participation by comparison with other groups.

ECDIS
0 = No discrimination
1 = Historical neglect/Remedial policies
Significant poverty and under representation in desirable occupations due to historical marginality, neglect, or restrictions. Public policies are designed to improve the group's material well being.
2 = Historical neglect/No remedial policies
Significant poverty and under representation due to historical marginality, neglect, or restrictions. No social practice of deliberate exclusion. Few or no public policies aim at improving the group's material well-being.
3 = Social exclusion/Neutral policies
Significant poverty and under representation due to prevailing social practice by dominant groups. Formal public policies toward the group are neutral or, if positive, inadequate to offset active and widespread discrimination.
4 = Restrictive policies
Public policies (formal exclusion And/or recurring repression) substantially restrict the group's economic opportunities by contrast with other groups (MAR, 2003: 39-40).

Appendix D: Interpolated Data

The data in Appendix D shows which data series had missing data filled in using interpolation. The blocks "Start" and "End" indicate the earliest data point used in the spline, and the last. The splines themselves are discussed in Chapters 2 and 3.

Country	Variable	Interpolation Method	Start	End
Yemen	Literacy	Cubic Spline	1975	2002
Somalia	Literacy	Cubic Spline	1977	2002
Sudan	Literacy	Cubic Spline	1975	2004
Djibouti	Literacy	Cubic Spline	1977	2003
Kenya	Literacy	Cubic Spline	1975	2002
Ethiopia	Literacy	Cubic Spline	1975	2002
Eritrea	Literacy	Linear Piecewise	1990	2005
Yemen	Education Gender Parity	Cubic Spline	1990	2004
Sudan	Education Gender Parity	Cubic Spline	1990	2004
Djibouti	Education Gender Parity	Cubic Spline	1989	2004
Kenya	Education Gender Parity	Cubic Spline	1990	2003
Ethiopia	Education Gender Parity	Cubic Spline	1990	2005
Eritrea	Education Gender Parity	Cubic Spline	1990	2004
Sudan	Agriculture as % of GNI	Cubic Spline	1975	2005
Yemen	Primary Commodity Ratio	Cubic Spline	1975	1990
Somalia	Primary Commodity Ratio	Cubic Spline	1975	1995
Sudan	Primary Commodity Ratio	Cubic Spline	1975	1995
Djibouti	Primary Commodity Ratio	Cubic Spline	1975	1995
Kenya	Primary Commodity Ratio	Cubic Spline	1975	1995
Ethiopia	Primary Commodity Ratio	Cubic Spline	1975	1995
Somalia	Forested Land	Cubic Spline	1990	2005
Sudan	Forested Land	Cubic Spline	1990	2005
Kenya	Forested Land	Cubic Spline	1990	2005
Ethiopia	Forested Land	Cubic Spline	1990	2005
Eritrea	Forested Land	Linear Piecewise	2000	2005
Djibouti	Infant Mortality Rate	Cubic Spline	1975	2006
Kenya	Infant Mortality Rate	Cubic Spline	1975	2006
Ethiopia	Infant Mortality Rate	Cubic Spline	1975	2006
Yemen	Undernourishment	Cubic Spline	1981	2004
Somalia	Undernourishment	Cubic Spline	1994	2006
Somalia	Undernourishment	Linear Piecewise	1989	1993
Sudan	Undernourishment	Cubic Spline	1981	2004
Djibouti	Undernourishment	Cubic Spline	1981	2004
Kenya	Undernourishment	Cubic Spline	1981	2004
Ethiopia	Undernourishment	Cubic Spline	1995	2004
Eritrea	Undernourishment	Cubic Spline	1995	2004

Appendix E: Extrapolated Data

This appendix shows what missing data was imputed via extrapolation, what type of model was used, and which parameters were used to define the model. The columns "Start" and "End" indicate the first and last year of the extrapolated missing data. Further discussion of extrapolation is found in Chapters 2 and 3.

Country	Variable	Type	Model	Start	End
Yemen	Adult Literacy	ARIMA	2,2,2	2003	2006
Somalia	Adult Literacy	ARIMA	2,2,2	2003	2006
Sudan	Adult Literacy	ARIMA	1,1,1	2005	2006
Djibouti	Adult Literacy	AR	1,0,0	2004	2006
Kenya	Adult Literacy	ARIMA	2,2,1	2003	2006
Ethiopia	Adult Literacy	ARIMA	2,2,1	2003	2006
Yemen	Agriculture as % of GNI	ARIMA	2,1,3	2004	2006
Sudan	Agriculture as % of GNI	ARIMA	2,2,2	2006	2006
Djibouti	Agriculture as % of GNI	ARIMA	2,2,2	2006	2006
Ethiopia	Agriculture as % of GNI	ARIMA	2,2,2	1975	1980
Eritrea	Agriculture as % of GNI	IMA	1,2	1990	1991
Yemen	Agriculture as a % GDP	ARIMA	2,1,1	2004	2006
Sudan	Agriculture as a % GDP	ARMA	3,1	2006	2006
Djibouti	Agriculture as a % GDP	ARIMA	2,1,1	2006	2006
Ethiopia	Agriculture as a % GDP	ARIMA	2,1,1	1975	1981
Eritrea	Agriculture as a % GDP	ARIMA	1,1,1	1990	1991
Djibouti	Aid as a % of GNI	ARIMA	2,2,1	2006	2006
Kenya	Aid as a % of GNI	ARIMA	1,1,1	2006	2006
Ethiopia	Aid as a % of GNI	ARIMA	1,1,1	2006	2006
Yemen	Aid Per Capita	ARIMA	3,1,1	2006	2006
Sudan	Aid Per Capita	ARIMA	3,2,1	2006	2006
Kenya	Aid Per Capita	ARMA	2,1	2006	2006
Ethiopia	Aid Per Capita	ARIMA	3,1,1	2006	2006
Yemen	Arable Land Per Capita	ARIMA	3,1,1	2004	2006
Somalia	Arable Land Per Capita	ARIMA	3,1,1	2004	2006
Sudan	Arable Land Per Capita	ARIMA	2,1,1	2004	2006
Kenya	Arable Land Per Capita	ARIMA	3,1,1	2004	2006
Ethiopia	Arable Land Per Capita	ARIMA	2,1,1	2004	2006
Eritrea	Arable Land Per Capita	ARIMA	3,1,3	1991	1992
Eritrea	Arable Land Per Capita	ARIMA	3,1,1	2004	2006
Yemen	Caloric Intake	ARMA	2,1	2004	2006
Somalia	Caloric Intake	ARMA	3,1	2004	2006
Kenya	Caloric Intake	ARMA	3,1	2004	2006
Ethiopia	Caloric Intake	ARIMA	3,1,1	2006	2006
Yemen	Education Gender Parity	ARIMA	1,1,1	2005	2006
Sudan	Education Gender Parity	ARIMA	1,1,1	2001	2006
Kenya	Education Gender Parity	ARIMA	2,2,2	2005	2006
Ethiopia	Education Gender Parity	ARIMA	1,1,1	2006	2006
Somalia	Forested Land	ARIMA	1,1,1	2006	2006
Sudan	Forested Land	ARIMA	1,1,1	2006	2006
Sudan	GDP Per Capita	ARIMA	2,2,2	2006	2006

Country	Variable	Type	Model	Start	End
Kenya	GDP Per Capita	ARI	1,1	2006	2006
Ethiopia	GDP Per Capita	ARIMA	2,1,1	1975	1980
Ethiopia	GDP Per Capita	ARIMA	2,1,1	2006	2006
Ethiopia	Life Expectancy	ARIMA	1,1,3	1975	1979
Djibouti	Military as a % of GDP	ARIMA	3,1,1	2003	2006
Kenya	Military as a % of GDP	ARIMA	1,1,1	2006	2006
Ethiopia	Military as a % of GDP	ARIMA	2,1,2	2006	2006
Yemen	Telephone Users	ARIMA	1,1,1	1975	1979
Yemen	Telephone Users	ARIMA	2,2,2	2005	2006
Somalia	Telephone Users	ARIMA	2,1,1	2005	2006
Sudan	Telephone Users	ARIMA	2,2,1	2005	2006
Djibouti	Telephone Users	ARIMA	2,2,1	2006	2006
Kenya	Telephone Users	ARIMA	2,2,1	2005	2006
Ethiopia	Telephone Users	ARIMA	2,2,1	2004	2006
Eritrea	Telephone Users	ARIMA	2,1,1	1990	1991
Eritrea	Telephone Users	ARIMA	3,1,1	2005	2006
Yemen	Trade Openness	ARIMA	1,1,1	2004	2006
Kenya	Trade Openness	ARMA	3,3	2004	2006
Ethiopia	Trade Openness	ARIMA	2,1,2	2004	2006
Yemen	Trade Ratio	ARIMA	2,1,2	2006	2006
Yemen	Undernourishment	ARIMA	2,1,1	2005	2006
Sudan	Undernourishment	ARIMA	3,3,1	2005	2006
Djibouti	Undernourishment	ARIMA	3,2,1	2005	2006
Djibouti	Undernourishment	ARIMA	3,2,1	1977	1980
Kenya	Undernourishment	ARIMA	2,1,1	2005	2006
Ethiopia	Undernourishment	ARIMA	1,1,1	2005	2006
Eritrea	Undernourishment	ARIMA	1,1,1	2006	2006
Eritrea	Undernourishment	ARIMA	2,1,1	1991	1995
Sudan	Youth Bulge	ARIMA	3,1,1	1975	1982
Djibouti	Youth Bulge	AR	3	1977	1983
Kenya	Youth Bulge	ARI	2,1	1975	1978
Somalia	Youth Bulge	ARIMA	1,1,1	2006	2006

Appendix F: Sample PCA Loadings Matrix

Variable	Comp 1	Comp 2	Com 3	Comp4	Comp5	Comp6	Comp7
Year	0.0242	0.0862	-0.1351	0.3475	0.2154	0.1976	0.061
Literacy	-0.0366	0.2147	-0.2316	0.0212	0.0093	0.2795	-0.0436
Gender.Parity	0.079	0.2726	-0.1299	0.0492	-0.0929	0.0324	0.196
PCE	-0.1795	-0.1514	-0.0379	-0.2565	-0.0309	-0.021	0.0534
Forested.Land	-0.0776	0.233	0.2674	0.0402	0.0843	0.037	-0.0185
LE	-0.0996	0.1194	-0.2152	-0.1432	0.1867	-0.1311	0.1964
IMR	0.0437	-0.2128	0.2194	-0.0057	-0.2451	0.0299	-0.1676
Bulge	-0.2196	-0.0443	-0.1269	-0.0721	0.1085	0.1579	0.1725
GDP98	0.2831	0.1218	-0.0719	0.0099	-0.0339	0.0499	-0.0506
open	0.25	0.0036	-0.1766	-0.0427	0.0571	-0.1545	0.0794
Population	-0.1577	0.1708	-0.0169	0.1715	-0.2182	-0.1288	-0.1627
Urban	0.2865	-0.0539	-0.0174	0.0012	0.0241	0.1889	-0.1094
Tel...100	0.0768	0.0254	-0.2554	0.0595	0.1782	0.153	-0.1716
Aid.as...GNI	0.0256	-0.1777	0.1364	0.0715	0.0154	0.2223	0.3792
Military.as...GDP	0.0908	0.0709	-0.0644	0.1132	0.2912	-0.278	0.2577
AG.as...GDP	-0.1974	-0.035	0.1969	0.1466	-0.2178	0.091	0.0745
Durability	-0.0146	-0.0259	-0.1155	-0.2807	-0.0813	0.1602	0.279
Trade.Ratio	-0.0082	0.0625	-0.246	-0.1948	-0.1262	0.0676	-0.1664
Aid.per.Cap	0.277	-0.0962	-0.0146	-0.0543	-0.0502	0.0661	0.0133
GDP.Growth	-0.0316	0.0704	0.0066	-0.0445	0.0614	-0.2882	-0.0082
Missing.Data	0.0053	-0.21	0.0675	-0.1975	-0.0124	-0.1553	-0.1935
Bad.Neighbors	0.007	0.2804	0.0487	-0.0528	-0.0928	0.1451	0.012
EF	0.0932	0.1464	0.1037	0.0924	-0.3189	0.199	0.1856
RF	0.0676	0.2868	-0.0484	0.0221	-0.2529	-0.1159	0.0387
LF	-0.1061	0.2637	-0.0769	0.0243	-0.2532	-0.1161	0.1047
Transition...88.	-0.0303	-0.0047	0.0071	0.1236	0.0912	-0.233	0.006
Anarchy...77.	-0.0348	-0.1123	0.1105	0.2306	0.003	0.2692	-0.0149
Full.Autocracy	0.0659	0.0328	0.0818	-0.3726	-0.1268	-0.0554	0.2914
Partial.Autocracy	-0.0569	-0.0114	-0.2118	0.0604	0.18	0.0402	-0.2353
Partial.Democracy.w.Factionalism	0.0073	0.0583	-0.0115	0.2263	-0.0419	-0.0404	-0.2116
Partial.Democracy.w.o.Factionalism	-0.0087	0.0298	-0.076	0.0485	-0.0159	0.086	-0.0876
Pol.Dis.1.	0.0424	0.1197	-0.0905	-0.1682	-0.1777	0.1541	-0.0586
Ec.Dis.1.	0.2425	0.136	0.0454	-0.0116	0.139	0.0468	0.0584
Years.since.last.conflict	0.1752	-0.0412	-0.1032	-0.1424	-0.1346	-0.0968	-0.1407
Change.in.Calories	0.0157	-0.0132	-0.0024	-0.0043	-0.0046	-0.0104	-0.1222
Change.in.IMR	0.0154	0.004	0.0125	0.0555	-0.1151	0.209	-0.0808
Pct.Paved	0.008	0.2667	0.2059	-0.0568	0.1702	0.0289	-0.0523
Km.Roads	-0.2182	-0.0486	-0.2545	-0.1166	-0.0307	0.0513	-0.0322
Calories	-0.0552	0.1385	-0.0991	-0.2817	0.1468	0.2332	-0.1899
Ed.as...GNI	-0.1362	-0.1625	-0.2111	-0.006	-0.2254	-0.0544	-0.1078
Water.Per.Capita	0.304	-0.0595	0.055	-0.0156	-0.0211	0.0063	-0.0024
Population.Density	-0.0876	0.113	-0.2397	0.2275	-0.1811	-0.1875	0.0949
Arable Land Per cap	-0.1309	0.2085	0.2282	-0.1363	0.0271	0.0148	-0.1566
Water/ AG interaction	0.3128	-0.0366	0.0018	-0.0275	-0.0249	-0.0339	-0.0368
Land Stress	-0.1099	0.2298	0.1722	-0.1836	0.1261	0.0071	-0.1587
Water / AG / Land	-0.0222	0.1727	0.2531	-0.1247	0.1871	-0.0066	-0.0459
Road per Cap	-0.0682	-0.0759	-0.0652	-0.0239	0.1088	0.223	0.1234
Relative GDP Per Cap	0.2904	0.1088	-0.0337	-0.0463	-0.0723	-0.0614	-0.0605

Appendix G: Sample Rotated PCA Loadings Matrix

Variable	Comp 1	Comp 2	Com 3	Comp 4	Comp 5	Comp 6	Comp 7
Year	0.0075	-0.0302	-0.0453	0.4414	0.1308	0.0039	-0.1456
Literacy	-0.012	0.1445	0.0249	0.1716	0.3439	0.1059	0.0323
Gender.Parity	0.0787	0.3195	-0.0041	0.1666	0.0458	-0.0541	0.0828
PCE	-0.1717	-0.0849	-0.0522	-0.1868	0.0223	-0.0235	0.2264
Forested.Land	-0.0657	0.0563	0.3356	0.089	-0.0893	0.0389	-0.0689
LE	-0.0983	0.0625	0.0009	0.0661	0.1187	-0.3385	0.1872
IMR	0.0429	-0.0846	-0.0456	-0.2309	-0.172	0.2886	-0.0848
Bulge	-0.2196	-0.059	-0.0463	0.1421	0.119	-0.0136	0.2227
GDP98	0.2924	0.0997	0.0095	0.0315	0.091	0.0174	-0.0369
open	0.2415	0.0338	-0.1235	-0.0029	0.0287	-0.2275	0.041
Population	-0.1547	0.2552	-0.0068	-0.0935	0.0022	0.0434	-0.2733
Urban	0.2957	-0.1042	-0.0124	0.0421	0.1053	0.1484	-0.014
Tel...100	0.0912	-0.1001	-0.0502	0.1199	0.3386	-0.0201	-0.0979
Aid.as...GNI	-0.0059	-0.0975	-0.0885	0.2828	-0.2343	0.1519	0.2722
Military.as...GDP	0.0594	-0.0118	-0.0121	0.2209	-0.1344	-0.4337	-0.0019
AG.as...GDP	-0.2111	0.0923	-0.0105	0.0407	-0.2063	0.252	-0.0281
Durability	-0.0067	0.0738	-0.0714	0.0016	0.0605	0.0342	0.4333
Trade.Ratio	0.0236	0.1052	-0.0632	-0.1923	0.305	0.0422	0.0567
Aid.per.Cap	0.2755	-0.0479	-0.0834	-0.0241	-0.0132	0.0707	0.0729
GDP.Growth	-0.0346	0.0443	0.053	-0.1149	-0.0724	-0.2566	-0.068
Missing.Data	0.0146	-0.1944	-0.012	-0.33	-0.0322	-0.0318	-0.0178
Bad.Neighbors	0.0305	0.2249	0.1926	0.0486	0.08	0.1143	0.0566
EF	0.0885	0.3041	-0.0098	0.1217	-0.1351	0.2868	0.0973
RF	0.0741	0.3945	0.0201	-0.0593	-0.0227	-0.0256	-0.0486
LF	-0.104	0.4015	-0.0129	-0.0371	-0.0254	-0.0411	-0.0006
Transition...88.	-0.0485	-0.0173	-0.0256	0.0145	-0.1078	-0.2023	-0.1512
Anarchy...77.	-0.047	-0.1225	-0.029	0.215	-0.0399	0.2812	-0.0836
Full.Autocracy	0.0735	0.1291	0.0634	-0.1571	-0.1476	-0.0647	0.4222
Partial.Autocracy	-0.0439	-0.1334	-0.0358	0.0258	0.2935	-0.0671	-0.1701
Partial.Democracy.w.Factionalism	0.0057	0.0447	-0.0154	0.0125	0.0397	0.0516	-0.3103
Partial.Democracy.w.o.Factionalism	-0.0015	0.0176	-0.0213	0.0277	0.1195	0.0681	-0.063
Pol.Dis.1.	0.0707	0.1627	0.0257	-0.1069	0.177	0.1496	0.1186
Ec.Dis.1.	0.2471	0.0046	0.1394	0.1346	0.0083	-0.0657	0.0374
Years.since.last.conflict	0.1878	0.0403	-0.091	-0.2435	0.0758	-0.0164	-0.0121
Change.in.Calories	0.0226	-0.0336	0.0118	-0.0697	0.0476	0.0225	-0.0781
Change.in.IMR	0.0243	0.0291	-0.0093	0.0351	0.0732	0.2409	-0.0286
Pct.Paved	0.0305	0.0227	0.3761	0.062	0.0072	-0.042	-0.0286
Km.Roads	-0.2058	0.0117	-0.1365	-0.0785	0.2356	-0.0192	0.0989
Calories	-0.005	-0.0731	0.221	-0.0698	0.3801	0.0527	0.1307
Ed.as...GNI	-0.1383	0.0487	-0.2816	-0.1964	0.0961	0.0638	-0.0473
Water.Per.Capita	0.2999	-0.0459	-0.0246	-0.018	-0.0737	0.0346	0.0095
Population.Density	-0.1101	0.3058	-0.2424	0.064	-0.0067	-0.1314	-0.1442
Arable Land Per cap	-0.0989	0.0385	0.3547	-0.1223	0.0167	0.0553	-0.0369
Water/ AG interaction	0.3117	-0.0209	-0.0399	-0.0515	-0.0306	-0.0075	-0.0157
Land Stress	-0.0732	-0.0007	0.38	-0.1042	0.0888	-0.028	-0.0084
Water / AG / Land	-0.0015	-0.0624	0.3712	-0.0074	-0.0444	-0.055	0.0189
Road per Cap	-0.0692	-0.1113	-0.0346	0.1715	0.0937	0.0719	0.174
Relative GDP Per Cap	0.2993	0.115	0.0167	-0.0754	0.0256	-0.0352	-0.0381

Appendix H: Sample Canonical Correlation Loadings Matrix

Variable	CC Score 1	CC Score 2	CC Score 3	CC Score 4
Year	-0.0253	-0.2225	0.1217	0.0265
Literacy	0.2155	-0.1044	0.1918	-0.0853
Gender.Parity	-0.1178	-0.1729	0.2625	-0.0042
PCE	0.2453	-0.0696	-0.2877	0.0254
Forested.Land	-0.1424	0.4769	0.5806	0.1015
LE	0.1768	-0.3392	0.134	0.1064
IMR	-0.126	0.3101	-0.314	-0.0317
Bulge	0.2803	0.0875	-0.1925	0.0887
GDP98	0.0479	-0.256	0.1559	-0.1296
open	0.0489	-0.555	-0.0019	-0.0664
Population	-0.4515	0.0645	0.1885	0.1885
Urban	0.3059	-0.1061	-0.0442	-0.1585
Tel...100	0.2462	-0.4308	-0.0806	-0.1624
Aid.as...GNI	0.127	0.1243	-0.2576	0.1698
Military.as...GDP	-0.3128	-0.4096	0.14	-0.2692
AG.as...GDP	-0.3134	0.4552	-0.2417	0.1864
Durability	0.2583	-0.1136	-0.1347	0.1385
Trade.Ratio	0.1686	-0.2137	-0.0214	-0.127
Aid.per.Cap	0.189	-0.2231	-0.0586	-0.1108
GDP.Growth	-0.1137	0.0227	-0.0185	0.2216
Missing.Data	0.0724	0.118	-0.1907	-0.2544
Bad.Neighbors	0.1397	0.0332	0.405	0.0717
EF	-0.075	0.1754	0.015	0.1094
LF	-0.3613	-0.0452	0.1808	0.1003
Transition...88.	-0.1211	-0.2546	0.1057	-0.078
Anarchy...77.	0.0434	0.1805	-0.1825	-0.2776
Full.Autocracy	0.0132	0.1401	-0.0747	0.0601
Partial.Autocracy	0.1073	-0.1091	-0.0502	-0.0744
Partial.Democracy.w.Factionalism	-0.0528	-0.0756	0.2867	0.3616
Pol.Dis.1.	-0.0807	0.2834	0.1294	-0.0266
Ec.Dis.1.	0.0476	-0.0362	0.494	-0.0948
Years.since.last.conflict	0.2269	-0.3065	-0.0988	-0.1004
Change.in.Calories	-0.0058	-0.1196	-0.0294	0.1486
Change.in.IMR	-0.0757	0.1144	0.0397	-0.0996
Pct.Paved	-0.0495	0.3729	0.6426	0.0179
Km.Roads	0.178	-0.0986	-0.1773	0.0471
Calories	0.6111	0.1545	0.231	0.0272
Ed.as...GNI	0.0654	-0.1544	-0.4613	0.0677
Water.Per.Capita	0.1772	-0.1565	-0.0355	-0.1096

Variable	CC Score 1	CC Score 2	CC Score 3	CC Score 4
Population.Density	-0.4629	-0.4759	-0.093	0.1347
Arable Land Per cap	-0.0424	0.553	0.3985	0.0674
Water/ AG interaction	0.1379	-0.2346	-0.0003	-0.1336
Land Stress	0.1286	0.4188	0.5442	0.0496
Water / AG / Land	0.1883	0.3257	0.4624	0.0826
Road per Cap	0.0362	0.0187	-0.054	0.0259
Relative GDP Per Cap	0.0188	-0.2208	0.1121	-0.1426
Somalia	0.2269	0.3462	-0.3306	0.1113
Kenya	0.2149	-0.1749	-0.1056	-0.018
Ethiopia	-0.7525	0.0149	-0.099	0.1244
Djibouti	0.1501	-0.2323	-0.065	-0.1583
Battle.Deaths.as...of.Pop	-0.5803	0.3694	-0.0959	-0.1497
Refugees.as...Pop	-0.6757	0.2819	-0.0601	-0.1344
Gen.Poli.per.Capita	0.0718	0.3608	0.5729	-0.056
Malnutrition	-0.6527	-0.4166	0.0759	0.0696

Appendix I: Normalization Transformations

The “~Normal” column indicates whether the final distribution of the data passes the Chi-Squared test with a p-value of .05 or greater. Note that some variables are transformed, but still do not pass the normality test. In these cases, the transformation was used to remove skewness.

Variable	Transformation	~Normal
Year	None	No
Literacy	None	Yes
Gender Parity	e^Xi	Yes
Primary Commodity Exports	None	No
Life Expectancy	None	No
Infant Mortality Rate	Xi^.5	Yes
Youth Bulge	Xi^2	No
Trade Openness	Xi^.5	Yes
Urban Population	ln (Xi)	No
Telephone Subscribers per 100 pop	ln (Xi)	No
Aid as a % of GNI	ln (Xi)	Yes
Military as a % of GDP	ln (Xi)	Yes
Agriculture as a % of GDP	None	Yes
Durability	Xi^.5	Yes
Trade Ratio	None	Yes
Foreign Aid per Cap	ln (Xi)	Yes
GDP Growth	None	Yes
Missing Data	Xi^.75	No
Bad Neighbors	Xi^.5	Yes
Ethnic Fractionalization	None	No
Religious Fractionalization	None	No
Linguistic Fractionalization	None	No
Transition Government	None	No
Full Autocracy	None	No
Partial Autocracy	None	No
Partial Democracy w/Factionalism	None	No
Political Discrimination	None	No
Economic Discrimination	None	No
Years since last conflict	None	No
Change in Calories	None	Yes
Change in Infant Mortality Rate	None	Yes
Pct Paved Roads	None	No
Calories	None	No
Education as a % of GNI	Xi^2	Yes
Water Per Capita	ln (Xi)	Yes
Population Density	ln (Xi)	Yes
Arable Land Per cap	Xi^2	Yes
Road per Cap	None	Yes
Relative GDP Per Cap	Xi^2	Yes

Appendix J: DA and Logistic Regression MATLAB Code

The MATLAB code shown in this section utilizes Fisher's Discriminant, and was tested on the Fisher's Iris data set to confirm the results conformed with other software results. The Discriminant code was also reviewed by the thesis reader, Dr. Kenneth Bauer. The logistic regression portion of the code uses the built in MATLAB mnrfit command. The rest of the code shown is here to automate the process of creating and defining groups by threshold, as well as recording the thousands of results.

```
function [Cp, APERgraph, DAchange, DAerrlog, DAtot, LogChangelog,
Logerrlog, Logtot, X0dist, X1dist, Xholddist] =  DAtest (X, new,
results, newy)
close all;
n = 1;
i = 1;
j = 1;
k = 1;
p = 1;
q = 1;
overunder = 0;
sizeX = size (X);
BestAPER = 1;
BestOverUnder = 0;
sizenew = size (new);
holdtruth = zeros (sizenew (1), 1);
APERgraph = [];
besttrngAPER = 1;
logtrnerrorcount = 0;
bestlogAPERtrngerr = 1;
bestlogAPERholderr = 1;
oldclassnew = zeros (sizenew (1), 2);
DAchange = [];
oldlogholdresult = zeros (sizenew (1), 1);
DAtot = zeros (sizenew (1), 1);
DAerrlog = [];
DAtestold = zeros (sizenew (1), 1);
LogChangelog = [];
Logtot = zeros (sizenew (1), 1);
Logerrlog = [];
X0dist = [];
X1dist = [];
Xholddist = [];
X0plot = [];
X1plot = [];
Xholdplot = [];
X0yearplot = [];
X1yearplot = [];
Xyearholdplot = [];

while overunder < 100001

% Create the two training matrices
while i < sizeX (1) + 1
    if results (i,1) <= overunder
```

```
        X0 (j,:) = X (i,:);
        if overunder == 0
            X0plot (j,1) = results (i,1);
            X0yearplot (j,1) = X (i,1);
        end
        j = j + 1;
    else
        X1 (k,:) = X (i,:);
        if overunder == 0
            X1plot (k,1) = results (i,1);
            X1yearplot (k,1) = X (i,1);
        end
        k = k+1;
    end

    i = i + 1;
end

% Create an over/under vector to build hold out confusion matrix
i = 1;

while i < sizenew (1) + 1
    if newy (i,1) <= overunder
        holdtruth (i,1) = 0;
    else
        holdtruth (i,1) = 1;
    end

    i = i + 1;
end

i = 1;

sizeX0 = size (X0);
sizeX1 = size (X1);

mu1 = mean (X0);
mu2 = mean (X1);

onevectX0 = ones (sizeX0(1),1);
onevectX1 = ones (sizeX1(1),1);
C1 = cov (X0);
C2 = cov (X1);

%Center B and O
XbarX0 = onevectX0'*X0/sizeX0(1);
XbarX1 = onevectX1'*X1/sizeX1(1);

Xdb = X0 - onevectX0*XbarX0;
Xdo = X1 - onevectX1*XbarX1;

%Find the pooled covariance

Cp = ((1/(sizeX0 (1) + sizeX1 (1) - 2))*(Xdb'*Xdb + Xdo'*Xdo));

% Singular Value decomposition is used to find the estimate inverse of
Cp
```

```
[U,S,V] = svd(Cp);
s       = diag(S);
e       = zeros(length(s),1);
ind     = s/max(abs(s)) >= eps;
e(ind)  = 1./s(ind);
E       = V*diag(e)*U';
%E is ~ Cp^-1

% Find d
d = mu1-mu2;

%Find maximum b value, also called Mahlanobis distance

maxb = d*E*d';

%Find T^2

T2 = ((sizeX0 (1) * sizeX1(1))/(sizeX0 (1) + sizeX1 (1)))*maxb;
Ft = (((sizeX0 (1) + sizeX1 (1)-sizeX0 (2)-1)/(sizeX0(2)*(sizeX0 (1) +
sizeX1 (1)-2))))*T2;

Fr = finv(.9, sizeX0(2), sizeX0 (1) * sizeX1(1) - sizeX0(2)-1);

%Find the mid-point
midpoint = .5* (mu1-mu2)*E*(mu1+mu2)';

%generate Confusion Matrix
i = 1;
bprime = (mu1-mu2)*E;

%Set the initial values for classifies correctly and classifies inc
N1c = 0;
N1cnot = 0;
N2c = 0;
N2cnot = 0;

% These are the prior probabilities

PP1 = (sizeX0(1))/(sizeX0(1)+sizeX1(1));
PP2 = (sizeX1(1))/(sizeX0(1)+sizeX1(1));

%Find confusion values for first sub type

while i < sizeX0(1)+1
      if bprime*X0(i,:)' <= midpoint
       N1cnot = N1cnot +1;
    else
        N1c = N1c +1;
    end

    i = i +1;
end

i = 1;

%Find confusion for second sub type
```

```
while i < sizeX1(1)+1
    if bprime*X1(i,:)' > midpoint
        N2cnot = N2cnot+1;
    else
        N2c = N2c +1;
    end

    i = i +1;
end

Cf = [N1c, N1cnot;N2cnot, N2c];

% Record the best training values for later use with the hold out data
APER = (N1cnot+N2cnot)/(sizeX0(1)+sizeX1(1));

% Record the scores when overunder = x
i = 1;

if overunder == 0

    while i < sizeX0 (1) + 1
        X0dist (i,1) = bprime*X0(i,:)';
        i = i + 1;
    end

    i = 1;

    while i < sizeX1 (1) + 1
        X1dist (i,1) = bprime*X1(i,:)';
        i = i + 1;
    end

end

if APER < besttrngAPER
    besttrngAPER = APER;
    bestCf = Cf;
    besttrngoverunder = overunder;
end

%Predict whether a data set indicates stable (0) or unstable (1)
i = 1;
classnew = zeros (sizenew (1), 1);


while i < sizenew (1) +1;
    if bprime*new(i,:)' > midpoint
        classnew (i,1) = i;
        classnew (i,2) = 0;
    else
        classnew (i,1) = i;
        classnew (i,2) = 1;
    end
    Xholddist (i,1) = bprime*new(i,:)';
    i = i + 1;
```

```
end

i = 1;

if overunder == 0
    while i < sizenew (1) + 1
        Xholddist (i,1) = bprime*new(i,:)';
        Xyearholdplot (i,1) = new (i,1);
        i = i + 1;
    end
end

i = 1;



% Display new confusion matrix for predicted data
i = 1;
N1c = 0;
N1cnot = 0;
N2c = 0;
N2cnot = 0;


while  i < sizenew (1) + 1
    if holdtruth (i) == 1
        if classnew (i,2) == 1
            N1c = N1c + 1;
        else
            N1cnot = N1cnot +1;
        end
    end

    if holdtruth (i) == 0
        if classnew (i,2) == 1
            N2cnot = N2cnot + 1;
        else
            N2c = N2c +1;
        end
    end

    i = i + 1;

end

%Make DA change and error matrices
i = 1;
flag = 0;
DAtest = classnew (:,2) - holdtruth;

if overunder == 0
    flag = 1;
end

while i < sizenew (1) + 1
    if DAtestold (i,1) ~= DAtest (i,1)
        flag = 1;
    end
    i = i + 1;
end

i = 1;

if flag == 1
    while i < sizenew (1) + 1
            DAchange (i,p) = classnew (i,2);
            DAerrlog (i,p) = DAtest (i,1);
            i = i + 1;
    end

    DAerrlog (sizenew (1) + 1, p) = overunder;
    DAchange (sizenew (1) + 1, p) = overunder;
    p = p+1;
    oldclassnew = classnew;
    DAtot = DAtot + DAtest;
    DAtestold = DAtest;
end

i = 1;
flag = 0;

%Record the results of the best outcomes

Ctest = [N1c, N1cnot;N2cnot, N2c];

APERtest = (N1cnot+N2cnot)/(sizenew (1));

if APERtest < BestAPER;
    BestAPER = APERtest;
    Ctest = [N1c, N1cnot;N2cnot, N2c];
    BestCtest = Ctest;
    BestOverUnder = overunder;
    bestestimate = classnew;
end

%Perform Logistic Regression for comparison
%First, make the ylog matrix consisting of 1's and 2's
i = 1;
ylog = [];

while i < sizeX (1) + 1
    if results (i,1) > overunder
        ylog (i,1) = 2;
    else
        ylog (i,1) = 1;
    end
    i= i+1;
end

i = 1;
B = mnrfit(X, ylog);
PHAT = mnrval(B,X);
logtest = [];
loghold = [];
logtrngconf = zeros (2,2);
```

```
logholdconf = zeros (2,2);
logtrnerrorcount = 0;
PHAThold = mnrval (B, new);
realylog = ylog - ones (sizeX (1),1);

%Build Predicted Outcome Matrices for both the training and Hold out
set

while i < sizeX (1) + 1
    if PHAT (i,2) > .5
        logtest (i,1) = 1;
    else
        logtest (i,1) = 0;
    end
    i= i+1;
end

i = 1;

while i < sizenew (1) + 1
    if PHAThold (i,2) > .5
        loghold (i,1) = 1;
    else
        loghold (i,1) = 0;
    end
    i= i+1;
end
i = 1;


%Find the APER and confusion matrices for training and hold out data
using
%Log Reg
logtestresult = logtest - realylog;

%Training Data First
while i < sizeX (1) + 1
    if logtestresult (i,1) == 0
        if logtest (i,1) == 1
           logtrngconf (1,1) = logtrngconf (1,1)+ 1;
        end

        if logtest (i,1) == 0
           logtrngconf (2,2) = logtrngconf (2,2)+ 1;
        end
    end

    if logtestresult (i,1) == -1
        logtrngconf (2,1) = logtrngconf (2,1)+ 1;
        logtrnerrorcount = logtrnerrorcount + 1;
    end

    if logtestresult (i,1) == 1
        logtrngconf (1,2) = logtrngconf (1,2)+ 1;
        logtrnerrorcount = logtrnerrorcount + 1;
    end

    i= i+1;
```

```
end
i = 1;

logAPERtrngerr = logtrnerrorcount / sizeX (1);

if logAPERtrngerr < bestlogAPERtrngerr
    bestlogAPERtrngerr = logAPERtrngerr;
    bestlogCf = logtrngconf;
    bestlogtrnoverunder = overunder;
end

%Now generate the APER and confusion matrix for the holdout data using
%logistic regression

logholdresult = loghold - holdtruth;
logtrnerrorcount = 0;

while i < sizenew (1) + 1
    if logholdresult (i,1) == 0
        if loghold (i,1) == 1
           logholdconf (1,1) = logholdconf (1,1)+ 1;
        end

        if loghold (i,1) == 0
           logtrngconf (2,2) = logtrngconf (2,2)+ 1;
        end
    end

    if logholdresult (i,1) == -1
        logholdconf (1,2) = logholdconf (1,2)+ 1;
        logtrnerrorcount = logtrnerrorcount + 1;
    end

    if logholdresult (i,1) == 1
        logholdconf (2,1) = logholdconf (2,1)+ 1;
        logtrnerrorcount = logtrnerrorcount + 1;
    end

    i= i+1;
end

%Make a Log Reg change matrix column every time the errors change
i = 1;
flag = 0;

while i < sizenew (1) + 1
    if oldlogholdresult (i,1) ~= logholdresult (i,1)
        flag = 1;
    end
    i = i + 1;
end

i = 1;

if overunder == 0
    flag = 1;
end
```

```
if flag == 1
    while i < sizenew (1) + 1
            LogChangelog (i,q) = loghold (i,1);
            Logerrlog (i,q) = logholdresult (i,1);
            i = i + 1;
    end

    LogChangelog (sizenew (1)+1, q) = overunder;
    Logerrlog (sizenew (1) + 1, q) = overunder;
    q = q+1;
    oldlogholdresult = logholdresult;
    Logtot = Logtot + logholdresult;
end

i = 1;

logAPERholderr = logtrnerrorcount / sizenew (1);

if logAPERholderr < bestlogAPERholderr
    bestlogAPERholderr = logAPERholderr;
    bestlogCtest = logholdconf;
    bestlogholdoverunder = overunder;
end


clear X1 X0 XbarB XbarO Xdb Xdo mu1 mu2 SizeB SizeO onevectB onevectO;

i = 1;
j = 1;
k = 1;
APERgraph (n,1) = overunder;
APERgraph (n,2) = APER;
APERgraph (n,3) = APERtest;
APERgraph (n,4) = logAPERtrngerr;
APERgraph (n,5) = logAPERholderr;
APERgraph (n,6) = Ctest (1,2);
APERgraph (n,7) = Ctest (2,1);
APERgraph (n,8) = logholdconf (1,2);
APERgraph (n,9) = logholdconf (2,1);

% Plots scores vs number of things measure (Battle Deaths, Genocide,
etc...

% if overunder == 0
%     figure(1)
%     plot (X0dist (:,1), X0plot (:,1), 'ro')
%     hold on
%     plot (X1dist (:,1), X1plot (:,1), 'bo')
%     hold on
%     plot (Xholddist (:,1), newy (:,1), 'go')
%
%     figure(2)
%     plot (X0dist (:,1), X0yearplot (:,1), 'ro')
%     hold on
%     plot (X1dist (:,1), X1yearplot (:,1), 'bo')
%     hold on
%     plot (Xholddist (:,1), Xyearholdplot (:,1), 'go')
%     hold on
% end


%overunder = overunder + 1;


if (overunder >= 0) && (overunder <100)
    overunder = overunder + 1;
end

if (overunder >= 100) && (overunder <1000)
    overunder = overunder + 10;
end

if (overunder >= 1000) && (overunder < 10000)
    overunder = overunder + 100;
end

if overunder >= 10000
    overunder = overunder + 1000;
end


n = n + 1;

end

bestlogtrnoverunder
bestlogholdoverunder
 bestlogCtest
 bestlogCf
 bestlogAPERtrngerr
 bestlogAPERholderr
 besttrngAPER
 bestCf
 BestCtest
 BestAPER
 BestOverUnder

end
```

Appendix K: Variables Used in Canonical Correlation and PCA

The 54 variables listed below were standardized, and then used to created the independent variable canonical correlation scores, and had also used to generate the initial PCA loadings matrix. The first 13 PCA loadings were then varimax rotated, and a set of 13 scores with 178 exemplars (each country from 1975 until 2002). These canonical and PCA scores were used to create the final models of each of the instability indicators: undernourisment, battle deaths per capita, refugees per capita, and genocide deaths.

Year	Partial Autocracy
Literacy	Partial Democracy w/Factionalism
Gender Parity	Political Discrimination
Primary Commodity Exports	Economic Discrimination
Forested Land	Years since last conflict
Life Expectancy	Change in Calories
Infant Mortality Rate	Change in IMR
Youth Bulge	Pct Paved
GDP per capita 98	Km Roads
Trade Openness	Calories
Population	Education as a % GNI
Urban	Water Per Capita
Telephone Subscribers per 100	Population Density
Foreign Aid as % GNI	Arable Land Per cap
Military as % of GDP	Water/ AG interaction
Agriculture as a % of GDP	Land Stress
Durability	Water / Agriculture / Land Interaction
Trade Ratio	Road per Cap
Foreign Aid per Cap	Relative GDP Per Cap
GDP Growth	Somalia
Missing Data	Kenya
Bad Neighbors	Ethiopia
EF	Djibouti
LF	4 Year Lagged Battle Deaths
Transition Government	4 Year Lagged Refugees
Anarchy	4 Year Lagged Genocide and Politicide Deaths
Full Autocracy	4 Year Lagged Malnutrition

Appendix L: Sample Canonical Correlation Scores from MI 1

These canonical correlation scores below were generated from the independent variables in Appendix K using the data from the first of the five multiply imputed data sets (MI 1).

Country	Year	Score 1	Score 2	Score 3	Score 4
Yemen	1975	0.8742	-0.1447	-0.4944	-0.6527
Yemen	1976	0.8198	-0.4555	-0.1056	0.2162
Yemen	1977	0.8573	-0.5251	-0.2491	1.4904
Yemen	1978	0.8222	-0.5882	0.03	0.0904
Yemen	1979	0.785	-0.968	-0.2779	1.2491
Yemen	1980	0.7446	-0.4616	-0.1891	0.4723
Yemen	1981	0.7379	-0.7856	0.1842	1.0419
Yemen	1982	0.7132	-0.6386	0.324	-0.2588
Yemen	1983	0.7787	-0.6628	0.2491	0.635
Yemen	1984	0.8037	-0.6721	-0.3239	0.2907
Yemen	1985	0.7127	-0.4248	0.0234	0.6273
Yemen	1986	0.7173	-0.523	-0.0773	-0.108
Yemen	1987	0.7306	-0.3917	0.1	0.5915
Yemen	1988	0.7397	-0.2755	0.1888	-0.183
Yemen	1989	0.7214	-0.531	-0.5385	-0.8695
Yemen	1990	0.8453	-0.8096	-0.1756	-0.7987
Yemen	1991	0.7444	-0.5598	-0.499	0.7804
Yemen	1992	0.7631	-0.5436	-0.69	0.5371
Yemen	1993	0.7529	-0.5047	-0.07	0.1681
Yemen	1994	0.7881	-0.8734	-0.0289	1.9376
Yemen	1995	0.8301	-0.7008	-0.2874	-0.1608
Yemen	1996	0.821	-0.5832	-0.1254	-0.6594
Yemen	1997	0.8407	-0.5248	-0.0566	0.6011
Yemen	1998	0.8478	-0.4222	-0.1426	0.0977
Somalia	1975	0.6576	-0.603	-0.5596	-0.0402
Somalia	1976	0.6876	-0.4188	-0.3412	0.2639
Somalia	1977	0.6081	-0.7346	-0.493	-0.6477
Somalia	1978	0.4892	-0.5014	-0.0545	0.5829
Somalia	1979	0.5247	-0.379	0.2541	-0.3528
Somalia	1980	0.7378	-0.4137	-0.1296	-0.1318
Somalia	1981	0.6889	-0.1182	-0.3966	-0.4939
Somalia	1982	0.6124	-0.0657	-0.6018	-0.2243
Somalia	1983	0.6627	0.427	0.3204	0.0506

Appendix M: Generalized Least Squares Model Battle Death Predictions

The results of various continuous regression models of battle deaths using raw data as part of a pilot study are shown below. Continuous models of battle deaths, refugees, and genocide were not part of the final recommended models.

Country	Year	Battle Deaths	OLS	Can Corr	polreg2/gauss	polreg3/gauss	Polreg2 / exp	Polreg3 / Cubic
Yemen	2003	0	450.5771	448.6811	1098.7059	297.143	1689.159	572.3397
Yemen	2004	0	2856.4047	2854.9799	356.5415	-169.2841	420.8826	-495.2589
Yemen	2005	0	898.9988	897.7045	5295.7329	7469.6751	5007.6734	4497.0632
Yemen	2006	0	2499.8924	2501.9975	24240.6829	24654.9173	23656.3397	25082.4835
Somalia	2003	0	212.1926	212.0109	455.5948	-2557.5519	382.5245	215.9903
Somalia	2004	0	-1734.4335	-1736.4063	6033.658	-3667.5005	5492.4051	6174.7332
Somalia	2005	0	-3152.2	-3153.8839	7890.8945	-4499.0217	7289.0634	7875.8056
Somalia	2006	547	-292.9191	-294.8239	6363.2133	-441.7778	5800.0149	6490.0289
Sudan	2003	3225	4186.0303	4181.7171	2897.2494	2404.8548	2383.5209	1768.6073
Sudan	2004	6569	6065.6459	6058.4292	5640.8656	4375.3495	5158.7447	5506.7249
Sudan	2005	1204	4164.7488	4157.1143	8131.3128	6234.4666	7492.5916	8200.7653
Sudan	2006	1002	5340.4629	5332.4808	19749.927	-138.4059	18669.0886	21536.8831
Djibouti	2003	0	8654.2601	8645.1123	3981.7154	4012.642	4172.8045	4107.9168
Djibouti	2004	0	9136.6584	9131.1086	3772.8423	3409.8132	4009.0417	3960.9683
Djibouti	2005	0	11588.5952	11579.77	5807.2251	4405.0248	5904.647	6404.1462
Djibouti	2006	0	10985.7681	10982.7107	1676.6655	914.2802	1500.8881	1448.5631
Kenya	2003	100	1518.2179	1519.4285	35.2892	-153.1825	14.8225	296.8486
Kenya	2004	52	-102.9236	-100.5088	-131.165	-3.5584	-255.4176	-368.5538
Kenya	2005	251	-1487.5769	-1485.879	6689.2627	7535.256	5732.6768	6480.507
Kenya	2006	567	261150.2682	261552.532	49940.8868	4781.4856	47358.7759	57218.4171
Ethiopia	2003	970	-3497.9586	-3501.066	6369.9844	6106.1041	8506.2527	4506.2289
Ethiopia	2004	936	2806.3613	2806.7765	6523.8632	6604.0947	8681.9833	5563.2227
Ethiopia	2005	773	2926.5821	2927.3815	6554.5047	4708.3301	7926.564	4621.1681
Ethiopia	2006	0	868.4189	870.7391	5821.6915	5297.0798	5769.955	6252.2793
Eritrea	2003	57	-1581.3616	-1580.3833	2182.4594	2843.2843	2963.8788	2467.7356
Eritrea	2004	0	2455.4234	2459.05	4750.3311	6426.2639	3908.232	5651.6528
Eritrea	2005	0	10063.6902	10064.2951	4218.289	4826.8667	4521.3128	10625.2949
Eritrea	2006	0	10671.6449	10672.1661	4679.8555	6461.3761	5053.7721	6313.9274
RMSE			51378.18427	51456.4474	12232.54093	6421.412682	11790.36072	13674.43847

Appendix N: Canonical Correlations Component Loadings

The loadings in this appendix represent the canonical correlation loadings of the independent variables in Appendix K generated using the first of the multiply imputed data sets. The data was standardized prior to using the canoncorr function in MATLAB. The loadings were found by finding the correlation between the input data, and the output canonical correlation scores.

	Canonical Variate Loading			
	1	2	3	4
Year	-0.0277	-0.2169	0.1302	0.039
Literacy	0.2147	-0.1048	0.1931	-0.088
Gender Parity	-0.1218	-0.1973	0.2134	0.0136
Primary Commodity Exports	0.1634	-0.168	-0.2699	0.0556
Forested Land	-0.1265	0.5005	0.5226	0.0824
Life Expectancy	0.1739	-0.3332	0.1474	0.1237
Infant Mortality Rate	-0.1233	0.3009	-0.3251	-0.0407
Youth Bulge	0.2793	0.1289	-0.1988	0.1176
GDP per capita 98	0.0332	-0.2666	0.1793	-0.1221
Trade Openness	0.0383	-0.5571	0.0182	-0.0235
Population	-0.45	0.0828	0.1916	0.182
Urban	0.3041	-0.117	-0.0447	-0.1525
Telephone Subscribers per 100	0.2399	-0.4406	-0.0696	-0.1287
Foreign Aid as % GNI	0.1337	0.0567	-0.257	0.1443
Military as % of GDP	-0.2247	-0.4385	0.1405	-0.0966
Agriculture as a % of GDP	-0.323	0.4527	-0.2428	0.1662
Durability	0.258	-0.1163	-0.1291	0.1478
Trade Ratio	0.2072	-0.246	-0.0291	-0.0971
Foreign Aid per Cap	0.1856	-0.2305	-0.0536	-0.0935
GDP Growth	-0.2001	0.0884	-0.0963	0.1669
Missing Data	0.0719	0.1025	-0.2001	-0.2554
Bad Neighbors	0.1421	0.0447	0.4048	0.0499
EF	-0.0721	0.1799	0.0112	0.0957
LF	-0.3458	-0.0146	0.1797	0.1187
Transition Government	-0.125	-0.2513	0.1136	-0.0601
Anarchy	-0.001	0.1095	-0.1693	-0.32
Full Autocracy	0.0404	0.1659	-0.0871	0.0812
Partial Autocracy	0.1055	-0.1142	-0.0482	-0.0652
Partial Democracy w/Factionalism	-0.0507	-0.0543	0.2968	0.3544
Political Discrimination	-0.0991	0.2629	0.1279	-0.0785
Economic Discrimination	0.0714	-0.0008	0.4877	-0.0857
Years since last conflict	0.2234	-0.3138	-0.0912	-0.0746
Change in Calories	-0.1304	-0.1474	0.048	-0.1504
Change in IMR	-0.1302	-0.1446	0.0485	-0.1543
Pct Paved	-0.1302	-0.1184	0.0856	-0.1526
Km Roads	0.213	-0.131	-0.2398	0.0225

	Canonical Variate Loading			
	1	2	3	4
Calories	0.6005	0.1874	0.1849	0.0501
Education as a % GNI	-0.0569	-0.2328	-0.2197	-0.087
Water Per Capita	0.1862	-0.15	-0.0371	-0.0842
Population Density	-0.1331	-0.1475	0.048	-0.1531
Arable Land Per cap	-0.136	-0.0019	0.1419	-0.1492
Water/ AG interaction	0.1483	-0.2243	-0.0007	-0.1037
Land Stress	0.1471	0.4358	0.5242	-0.0004
Water / Agriculture / Land Interaction	0.2074	0.3502	0.4475	0.0501
Road per Cap	0.2215	-0.2296	-0.168	-0.0974
Relative GDP Per Cap	0.0177	-0.2194	0.1224	-0.1056
Somalia	0.2318	0.3357	-0.3415	0.0935
Kenya	0.2127	-0.1811	-0.1007	-0.0035
Ethiopia	-0.1306	-0.1446	0.0485	-0.1542
Djibouti	-0.1302	-0.1446	0.0485	-0.1543
4 Year Lagged Battle Deaths	-0.5622	0.3918	-0.1172	-0.1427
4 Year Lagged Refugees	-0.6659	0.294	-0.0732	-0.1321
4 Year Lagged Genocide and Politicide Deaths	0.078	0.3723	0.5591	-0.1081
4 Year Lagged Malnutrition	-0.6051	-0.3453	0.0765	0.1596

Appendix O: PCA Loadings for MI 1

The PCA loadings shown in Appendix O were generated using the variables shown below, and 178 exemplars from 7 countries covering the years 1975-2002. These are the same variables listed in Appendix K. The data used to generate the matrix below comes from the first of the five multiply imputed data sets (MI 1). The data was standardized prior to principal component analysis being applied, and has not been rotated. Only the first 13 (retained) principal components are shown. The loadings were generated using the built in MATLAB function "princomp"

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
Year	0.0172	0.075	-0.044	0.1258	0.2488	0.3659	-0.116	-0.214	-0.122	0.0061	-0.021	0.1229	-0.046
Literacy	-0.0491	0.1734	-0.251	0.0088	-0.04	0.2331	-0.055	4E-04	-0.039	0.1116	0.009	0.1367	0.015
Gender Parity	0.0402	0.2472	-0.163	0.1169	-0.011	0.0766	0.2162	0.027	0.077	0.0278	0.003	0.0523	-0.081
Primary Commodity Exports	-0.1158	-0.1784	-0.114	-0.125	0.1131	-0.157	0.19	0.036	0.041	-0.065	-0.107	-0.058	-0.018
Forested Land	-0.1466	0.2442	0.1708	-0.096	0.035	0.0139	0.0111	-0.029	-0.007	-0.017	-1E-04	-4E-04	0.011
Life Expectancy	-0.0777	0.0615	-0.27	-0.043	0.1512	-0.049	0.1451	-0.021	0.118	-0.171	0.036	-0.098	-0.069
Infant Mortality Rate	0.0518	-0.1679	0.254	0.0174	-0.224	-0.069	-0.062	0.012	-0.021	-0.031	0.009	-0.033	0.048
Youth Bulge	-0.1884	-0.1368	-0.141	-0.127	0.0528	0.1052	0.0968	-0.093	-0.058	0.0352	-0.128	-0.064	0.021
GDP per capita 98	0.2469	0.1685	-0.045	0.0651	-0.051	0.0357	-0.062	-0.054	-0.068	0.0074	0.025	0.0576	-0.007
Trade Openness	0.2549	0.0309	-0.124	0.0549	0.0835	-0.099	0.0467	0.044	0.016	-0.001	-0.055	-0.041	-0.018
Population	-0.189	0.1215	0.0057	0.2488	-0.092	-0.032	-0.157	0.022	-0.105	-0.124	0.034	-0.024	-0.018
Urban	0.2801	0.0211	0.0318	-0.047	-0.071	0.1142	-0.094	-0.042	-0.021	-0.009	0.057	0.0336	-0.008
Telephone Subscribers per 100	0.1788	0.0008	-0.223	0.045	0.09	0.0908	-0.18	-0.113	0.069	0.0533	0.107	0.0423	0.073
Foreign Aid as % GNI	0.099	-0.1279	0.1607	-0.067	0.0025	0.2392	0.2788	-0.062	-0.013	-0.169	-0.014	-0.144	0.084
Military as % of GDP	0.0813	0.0003	-0.015	0.0307	0.3801	-0.183	0.1746	-0.117	-0.05	0.1064	-0.026	-0.076	0.05
Agriculture as a % of GDP	-0.2024	-0.0579	0.1878	0.0916	-0.13	0.126	0.1332	0.033	0.038	-0.096	0.035	-0.05	0.062
Durability	0.0212	-0.0619	-0.187	-0.081	-0.188	-0.002	0.1317	-0.295	-0.361	-0.016	-0.029	0.0314	0.035
Trade Ratio	0.0107	0.0007	-0.278	-0.001	-0.174	-0.15	-0.072	0.097	-0.065	0.1774	-0.114	0.1081	-0.027
Foreign Aid per Cap	0.2757	-0.017	0.0249	-0.017	-0.082	0.0407	-0.026	0.011	-0.077	-0.001	0.01	-0.037	0.013
GDP Growth	-0.054	0.0366	0.0681	0.0403	0.1309	0.0379	0.1354	0.151	0.379	-0.01	-0.256	0.2256	0.111
Missing Data	0.0435	-0.183	0.0413	-0.101	-0.064	-0.284	-0.072	0.161	0.131	0.1644	-0.191	0.0866	-0.047
Bad Neighbors	-0.0437	0.2644	-0.045	-0.06	-0.117	0.0727	0.0699	0.193	-0.122	0.0345	0.071	-0.07	0.085
EF	0.0402	0.1691	0.087	0.069	-0.22	0.2287	0.2959	0.173	0.009	-0.072	0.027	-0.033	-0.013
LF	-0.1434	0.2065	-0.119	0.1808	-0.105	-0.033	0.1555	0.148	-0.006	0.0144	-0.018	0.0206	0.014
Transition Government	-0.0278	-0.009	0.0138	0.0812	0.217	-0.005	-0.062	0.229	-0.234	-0.045	0.301	-0.168	0.602
Anarchy	-0.0316	-0.0891	0.1376	-0.001	0.0231	0.3131	-0.026	0.188	0.012	0.3563	-0.026	0.3766	-0.186
Full Autocracy	0.069	0.0418	-0.044	-0.105	-0.24	-0.261	0.2991	-0.285	-0.068	0.0081	-0.035	0.0692	-0.046
Partial Autocracy	-0.0209	-0.0492	-0.122	0.0348	0.1513	0.1097	-0.242	-0.095	0.508	-0.029	0.177	-0.118	-0.016
Partial Democracy w/Factions	-0.0463	0.0808	0.0452	0.0785	0.0391	0.0481	-0.231	0.192	-0.158	-0.335	-0.43	-0.245	-0.361
Political Discrimination	0.0414	0.1245	-0.081	0.0573	-0.257	0.0755	0.0328	-0.246	0.391	-0.129	0.022	-0.23	0.148
Economic Discrimination	0.1973	0.2007	0.0284	-0.045	0.0549	0.0188	-0.037	-0.16	-0.037	0.0045	-0.151	-0.011	-0.014
Years since last conflict	0.1952	-0.0157	-0.069	-0.007	-0.118	-0.067	0.0276	0.324	0.212	-0.065	-0.005	-0.074	0.061
Change in Calories	0.0215	0.0005	0.0017	-0.008	0.0065	-0.044	-0.054	0.004	0.021	-0.504	-0.212	0.6043	0.341
Change in IMR	0.0001	0.0146	0.0294	0.038	-0.096	0.0996	-0.087	0.011	-0.027	0.4116	-0.499	-0.211	0.476
Pct Paved	-0.0628	0.2857	0.1266	-0.153	0.0376	-0.077	-0.036	-0.065	-0.015	-0.009	0.006	-0.005	0.03
Km Roads	-0.1563	-0.152	-0.258	-0.028	-0.068	0.0054	0.0034	0.022	0.058	0.0072	-0.022	-0.036	0.04

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
Education as a % GNI	-0.0574	-0.2678	-0.141	0.0917	-0.142	-0.056	-0.13	0.054	-0.024	0.0033	0.05	-0.035	-0.004
Water Per Capita	0.2821	0.0277	0.0894	-0.029	-0.033	0.0153	0.049	0.093	0.041	-0.006	0.032	-0.034	-0.019
Population Density	-0.0912	0.0296	-0.174	0.3503	0.0479	0.0202	0.0319	0.045	-0.097	-0.055	-0.038	-0.003	-0.017
Arable Land Per cap	-0.1781	0.2007	0.1041	-0.15	-0.074	-0.134	-0.063	0.03	0.054	0.0477	0.109	0.0479	0.033
Water/ AG interaction	0.2911	0.0457	0.0445	0.0037	-0.032	-0.021	0.0017	0.075	0.02	0.0122	0.035	-0.015	-0.029
Land Stress	-0.151	0.2176	0.0371	-0.213	-0.01	-0.146	-0.095	0.044	0.023	0.064	0.077	0.0335	4E-04
Water / Agriculture / Land	-0.0695	0.195	0.1344	-0.248	0.0688	-0.1	0.0207	0.122	0.009	0.0392	0.134	0.0192	-0.009
Road per Cap	0.1217	-0.14	-0.05	-0.097	-0.064	0.0342	0.0021	-0.02	-0.033	-0.174	-0.061	0.1012	0.138
Relative GDP Per Cap	0.2538	0.1631	-0.023	0.0538	-0.062	-0.066	-0.01	0.06	0.024	0.0388	0.018	0.0344	-0.011
Somalia	-0.0334	-0.1758	0.18	-0.127	-0.059	0.2887	0.1855	-4E-04	0.049	-0.098	-0.008	-0.09	-0.024
Kenya	-0.0614	0.0567	-0.277	0.0124	-0.146	0.1264	0.2123	0.197	0.057	0.0612	-0.063	-0.004	0.019
Ethiopia	-0.1002	0.0091	0.1065	0.3582	-0.114	-0.126	-0.074	-0.014	-0.073	-0.092	0.058	0.0181	-0.021
Djibouti	0.0325	0.059	0.0305	0.0654	0.3834	-0.079	0.3275	-0.026	-0.01	0.0548	-0.142	0.0111	-0.003
4 Year Lagged Battle Deaths	-0.0871	0.0149	0.1231	0.2173	-0.132	-0.083	0.0281	-0.284	0.17	0.1034	-0.008	-0.007	0.036
4 Year Lagged Refugees	-0.0889	0.0301	0.1573	0.2338	-0.064	-0.054	0.0501	-0.257	0.092	0.1623	0.021	0.205	0.055
4 Year Lagged Genocide Deaths	-0.043	0.1935	0.0691	-0.124	0.0128	0.0557	-0.196	-0.163	0.044	-0.056	-0.346	-0.189	0.053
4 Year Lagged Malnutrition	0.1187	0.0557	0.0699	0.3152	0.0756	-0.237	0.0582	0.055	0.047	-0.058	-0.082	-0.058	0.019

A suggested interpretation of each of these principal components is listed in Table 4-2.

They are not shown here in the interest of space.

Appendix P: PCA OLS Model Results

The results shown here represent model parameters of continuous models using 54 variables, 13 principlal component scores, and OLS regression.

	Battle Deaths	Refugees	Genocide	Malnutrition
R square	0.3316	0.6722	0.4763	0.8315
R square Adj	0.2677	0.6409	0.4222	0.8154
RMSE	4855.78	262112	36286.5	6.65
Var 1 Coefficient	-370.50606	-38620.1	-1986.51	1.3870803
Var 2 Coefficient	170.338311	11956.08	8215.551	0.755095
Var 3 Coefficient	531.729788	74605.69	3078.62	0.8267673
Var 4 Coefficient	887.19991	74449.9	-6565.31	4.6394338
Var 5 Coefficient	-104.3907	-40749.2	-146.807	1.2588538
Var 6 Coefficient	-408.81465	-39011.8	-892.594	-3.8012375
Var 7 Coefficient	456.982967	23937.14	-6418.14	1.1958745
Var 8 Coefficient	-570.99917	-98090.2	-4623.19	1.6588628
Var 9 Coefficient	64.0519084	76231.83	-42.4988	0.3639824
Var 10 Coefficient	-851.55989	58038.35	1124.969	0.8342598
Var 11 Coefficient	-501.20001	-15400.7	-2813.61	0.0151556
Var 13 Coefficient	-184.5208	25770.03	-7265.47	-0.0851376
Var 13 Coefficient	-109.49139	24401.62	302.6749	0.5159077
t-ratio of Variable 1	-3.0835822	-5.95451	-2.2124	8.426912
t-ratio of Variable 2	1.20868847	1.571676	7.801033	3.9112024
t-ratio of Variable 3	3.63680468	9.453077	2.817724	4.127803
t-ratio of Variable 4	4.98767808	7.753777	-4.93907	19.039194
t-ratio of Variable 5	-0.5130925	-3.71044	-0.09656	4.5166433
t-ratio of Variable 6	-1.82088	-3.21902	-0.53201	-12.359091
t-ratio of Variable 7	1.86799036	1.812671	-3.51073	3.5683454
t-ratio of Variable 8	-1.8287069	-5.81978	-1.98136	3.8781603
t-ratio of Variable 9	0.19056627	4.201676	-0.01692	0.7904983
t-ratio of Variable 10	-2.4099322	3.042825	0.426033	1.7234457
t-ratio of Variable 11	-1.3695857	-0.77964	-1.02886	0.0302314
t-ratio of Variable 12	-0.4651008	1.203342	-2.45064	-0.1566498
t-ratio of Variable 13	-0.2694877	1.112628	0.099689	0.9269103
Prob Variable 1 > |t|	0.00247732	2.11E-08	0.028607	4.51E-14
Prob Variable 2 > |t|	0.22887984	0.118349	1.45E-12	0.0001444
Prob Variable 3 > |t|	0.00039071	1.30E-16	0.005558	6.35E-05
Prob Variable 4 > |t|	1.83E-06	1.88E-12	2.26E-06	3.49E-40
Prob Variable 5 > |t|	0.60871907	0.000301	0.923219	1.35E-05
Prob Variable 6 > |t|	0.07082329	0.001609	0.595585	5.44E-24
Prob Variable 7 > |t|	0.06391526	0.072088	0.000607	0.0004969
Prob Variable 8 > |t|	0.06963445	4.03E-08	0.049567	0.0001632
Prob Variable 9 > |t|	0.84914959	4.77E-05	0.986525	0.4306131
Prob Variable 10 > |t|	0.01729302	0.002814	0.670757	0.0870806
Prob Variable 11 > |t|	0.17307344	0.43696	0.305373	0.9759268
Prob Variable 12 > |t|	0.64260314	0.230935	0.01553	0.8757532
Prob Variable 13 > |t|	0.78796284	0.267831	0.920738	0.3556151

Appendix Q: DACE Model Results

The tables below show the best continuous model results for the instability indicators. Multiply Imputed data set 2 (MI 2) was used to generate these numbers, and the MATLAB DACE program to build the individual models. The PCA and canonical correlation scores were generated using the 54 variables in Appendix K.

Battle Deaths

	PCA	CC	Raw Data
OLS Rsquare	0.4602	0.4927	0.4927
OLS RMSE	4361.1	5293.2	5293.2
Best GLS Model Polynomial	1	1	N/A
Best GLS Model Correlation	Gaussian	Spline	N/A
Best GLS Model Rsquare	0.4699	0.4542	N/A
Best GLS Model RMSE	3940.5	3563.7	N/A

Refugees

	PCA	CC	Raw Data
OLS Rsquare	0.6174	0.5234	0.5234
OLS RMSE	275905.4	270994.0	271002.5
Best GLS Model Polynomial	1	1	N/A
Best GLS Model Correlation	Gaussian	Linear	N/A
Best GLS Model Rsquare	0.6523	0.3760	N/A
Best GLS Model RMSE	244731.0	247407.7	N/A

Genocide

	PCA	CC	Raw Data
OLS Rsquare	0.5253	0.4656	0.2682
OLS RMSE	26436.2	66232.8	265732.9
Best GLS Model Polynomial	1	1	N/A
Best GLS Model Correlation	Exponential	Spline	N/A
Best GLS Model Rsquare	0.5656	0.3867	N/A
Best GLS Model RMSE	25815.2	56843.0	N/A

Malnutrition

	PCA	CC	Raw Data
OLS Rsquare	0.8087	0.9202	0.9202
OLS RMSE	7.0327	4.8367	4.8374
Best GLS Model Polynomial	1	1	N/A
Best GLS Model Correlation	Gaussian	Exponential	N/A
Best GLS Model Rsquare	0.8422	0.9204	N/A
Best GLS Model RMSE	6.6893	4.8304	N/A

Appendix R: Individual Country Predictions

The predictions shown below were generated by building OLS models using canonical correlation scores. The scores, and the OLS models, were generated using only exemplars from a single country at a time, rather than aggregating the data into a single data set,as was done everywhere else in the research.

		Battle Deaths		Refugees		Genocide Deaths		Malnutrition	
Country	Year	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Yemen	2003	0	1254.768	550	-61268.863	0	N/A	37	37.797
Yemen	2004	0	5711.511	526	64027.653	0	N/A	38	41.0002
Yemen	2005	0	5943.013	395	94979.444	0	N/A	30.4838	41.6217
Yemen	2006	0	7697.604	573	92189.485	0	N/A	36.6368	43.663
Somalia	2003	0	24751.95	315114	840835.53	0	2348.869	16.1	21.1839
Somalia	2004	0	111892.6	306200	1481753.4	0	-25403.3	19.7	32.7654
Somalia	2005	0	196360	314066	2850882.4	0	-14872.2	12.7	26.5539
Somalia	2006	547	210897.4	388046	3780116.1	0	8061.195	18.6	23.9713
Sudan	2003	3225	-10510.1	580727	471353.63	96000	-64710.7	27	31.8646
Sudan	2004	6569	-18078.11	696657	571361.92	192000	-417826	26	29.9466
Sudan	2005	1204	-17697.57	655732	4825.3956	48000	16443.43	29.2417	38.4199
Sudan	2006	1002	-13291.01	645093	151602.52	48000	-47945	32.3014	27.8723
Djibouti	2003	0	25.4938	78	1495.1208	0	N/A	26	25.8675
Djibouti	2004	0	14.0676	73	-11084.796	0	N/A	24	22.108
Djibouti	2005	0	22.2872	77	-21286.592	0	N/A	25.2437	16.526
Djibouti	2006	0	-183.511	105	-53304.71	0	N/A	27.5347	8.8973
Kenya	2003	100	-6606.034	883	-4458.2845	0	N/A	31	27.2425
Kenya	2004	52	4892.203	935	-12051.367	0	N/A	31	20.5406
Kenya	2005	251	-2093.936	1186	-41057.472	0	N/A	28.2095	16.8291
Kenya	2006	567	-15847.82	1753	-4657.9218	0	N/A	27.6009	27.7872
Ethiopia	2003	970	-63556.44	43676	-152264.14	0	-396.681	46	43.3017
Ethiopia	2004	936	-50216.79	44219	3426269.4	0	241.1668	46	60.4557
Ethiopia	2005	773	-127617.9	46824	7441664.8	0	585.2387	48.6663	76.0819
Ethiopia	2006	0	-127721	65361	7171569.3	0	479.4786	48.366	66.2827
Eritrea	2003	57	148569.3	116964	722124.22	0	N/A	73	39.0678
Eritrea	2004	0	-1594122	121937	-2261749.5	0	N/A	75	464.0699
Eritrea	2005	0	-2470262	130544	-3825423.6	0	N/A	72.2599	675.7927
Eritrea	2006	0	-1572842	168682	-2313586.1	0	N/A	71.7162	456.582
RMSE		658572.2766		2517962.018		202201.5993		160.2417213	

Appendix S: Discrete Model Results

Note: The highlighted totals indicate the best overall model in terms of false negatives, false positives, total error, and apparent error. The number in "Data Type" indicates the number of variables used in the particular model. The green blocks indicate the lowest.

Battle Deaths Discrete Model Results:

	Data Type	38 Raw, No Transform		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	2406	6573	8979	0.2565
MI2	2418	6951	9369	0.2676
MI3	2779	6624	9403	0.2686
MI4	2682	7345	10027	0.2864
MI5	2485	4691	7176	0.2050
Total	12770	32184	44954	0.2568

	Data Type	38, Normalized		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1496	8750	10246	0.2927
MI2	2585	7907	10492	0.2997
MI3	1717	9066	10783	0.3080
MI4	1492	8813	10305	0.2944
MI5	1507	9000	10507	0.3002
Total	8797	43536	52333	0.2990

	Data Type	38 Raw, No Transform		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	1145	14550	15695	0.4484
MI2	2171	16192	18363	0.5246
MI3	1413	15136	16549	0.4728
MI4	1793	15396	17189	0.4911
MI5	1059	16765	17824	0.5092
Total	7581	78039	85620	0.4892

	Data Type	38 Raw, Normalized		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	4656	6812	11468	0.3276
MI2	6395	7807	14202	0.4057
MI3	6565	6195	12760	0.3645
MI4	5802	6744	12546	0.3584
MI5	4718	10158	14876	0.4250
Total	28136	37716	65852	0.3762

	Data Type	13 PCA Scores		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1791	4663	6454	0.1844
MI2	1546	5619	7165	0.2047
MI3	1571	4456	6027	0.1722
MI4	1784	6668	8452	0.2414
MI5	1546	4128	5674	0.1621
Total	8238	25534	33772	0.1929

	Data Type	13 PCA Scores		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	2169	4349	6518	0.1862
MI2	2264	6213	8477	0.2422
MI3	3148	4764	7912	0.2260
MI4	2199	5802	8001	0.2286
MI5	2925	3628	6553	0.1872
Total	12705	24756	37461	0.2140

	Data Type	PCA Scores w/data Per Capita		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1024	6793	7817	0.2233
MI2	964	8207	9171	0.2620
MI3	984	7264	8248	0.2356
MI4	1024	9413	10437	0.2982
MI5	1025	6871	7896	0.2256
Total	5021	38548	43569	0.24896

	Data Type	PCA Scores w/data Per Capita		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	1853	3585	5438	0.1553
MI2	1832	4259	6091	0.1740
MI3	1875	3785	5660	0.1617
MI4	1948	5123	7071	0.2020
MI5	1853	3749	5602	0.1600
Total	9361	20501	29862	0.1706

Refugees Discrete Model Results

	Data Type	38 Raw, No Transform		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1384	93	1477	0.1433
MI2	1081	154	1235	0.1198
MI3	832	192	1024	0.0993
MI4	968	287	1255	0.1217
MI5	1172	230	1402	0.1360
Total	5437	956	6393	0.1240

	Data Type	38, Normalized		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	241	743	984	0.0954
MI2	845	524	1369	0.1328
MI3	233	864	1097	0.1064
MI4	572	1070	1642	0.1593
MI5	809	780	1589	0.1542
Total	2700	3981	6681	0.1296

	Data Type	38 Raw, No Transform		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	240	752	992	0.0962
MI2	989	453	1442	0.1399
MI3	51	1219	1270	0.1232
MI4	252	1375	1627	0.1578
MI5	1037	551	1588	0.1541
Total	2569	4350	6919	0.1342

	Data Type	38 Raw, Normalized		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	321	865	1186	0.1151
MI2	1084	575	1659	0.1610
MI3	254	932	1186	0.1151
MI4	526	1317	1843	0.1788
MI5	826	864	1690	0.1640
Total	3011	4553	7564	0.1468

	Data Type	13 PCA Scores		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	180	673	853	0.0827
MI2	140	808	948	0.0920
MI3	189	607	796	0.0772
MI4	265	730	995	0.0965
MI5	177	617	794	0.0770
Total	951	3435	4386	0.0851

	Data Type	13 PCA Scores		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	135	595	730	0.0708
MI2	182	242	424	0.0411
MI3	138	510	648	0.0628
MI4	140	410	550	0.0533
MI5	268	292	560	0.0543
Total	863	2049	2912	0.0565

	Data Type	PCA Scores w/data Per Capita		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	190	833	1023	0.0992
MI2	163	826	989	0.0959
MI3	116	815	931	0.0903
MI4	245	841	1086	0.1053
MI5	115	823	938	0.0910
Total	829	4138	4967	0.0964

	Data Type	PCA Scores w/data Per Capita		
	Method	Logistic Regression		
	False -	False +	Total Errors	APER
MI1	71	560	631	0.0612
MI2	76	451	527	0.0511
MI3	94	479	573	0.0556
MI4	78	550	628	0.0609
MI5	113	407	520	0.0504
Total	432	2447	2879	0.0558

Genocide / Politicide Discrete Model Results

	Data Type	38 Raw, No Transform		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1025	576	1601	0.1425
MI2	1076	0	1076	0.0958
MI3	986	34	1020	0.0908
MI4	938	576	1514	0.1348
MI5	983	34	1017	0.0905
Total	5008	1220	6228	0.1109

	Data Type	38, Normalized		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	1001	0	1001	0.0891
MI2	1085	34	1119	0.0996
MI3	1001	17	1018	0.0906
MI4	1004	0	1004	0.0894
MI5	995	0	995	0.0886
Total	5086	51	5137	0.0915

	Data Type	38 Raw, No Transform		
	Method	Log Reg		
	False -	False +	Total Errors	APER
MI1	1118	1782	2900	0.2582
MI2	1121	1169	2290	0.2039
MI3	1112	955	2067	0.1840
MI4	1121	1587	2708	0.2411
MI5	1085	2203	3288	0.2928
Total	5557	7696	13253	0.2360

	Data Type	38 Normalized		
	Method	Log Reg		
	False -	False +	Total Errors	APER
MI1	1133	82	1215	0.1082
MI2	1097	40	1137	0.1012
MI3	1085	969	2054	0.1829
MI4	1169	51	1220	0.1086
MI5	1133	226	1359	0.1210
Total	5617	1368	6985	0.1244

	Data Type	PCA Scores		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	0	435	435	0.0387
MI2	0	435	435	0.0387
MI3	0	435	435	0.0387
MI4	0	435	435	0.0387
MI5	0	435	435	0.0387
Total	0	2175	2175	0.0387

	Data Type	PCA Scores		
	Method	Log Reg		
	False -	False +	Total Errors	APER
MI1	995	334	1329	0.1183
MI2	977	17	994	0.0885
MI3	977	192	1169	0.1041
MI4	977	641	1618	0.1441
MI5	977	353	1330	0.1184
Total	4903	1537	6440	0.1147

	Data Type	PCA Scores w/data Per Capita		
	Method	DA		
	False -	False +	Total Errors	APER
MI1	0	710	710	0.0632
MI2	0	818	818	0.0728
MI3	0	710	710	0.0632
MI4	3	742	745	0.0663
MI5	0	874	874	0.0778
Total	3	3854	3857	0.0687

	Data Type	PCA Scores w/data Per Capita		
	Method	Log Reg		
	False -	False +	Total Errors	APER
MI1	358	1	359	0.0319
MI2	352	141	493	0.0439
MI3	358	58	416	0.0370
MI4	358	366	724	0.0644
MI5	349	173	522	0.0464
Total	1775	739	2514	0.0447

The green blocks indicate the lowest value for each category. For instance, the 0 false positives with PCA scores and DA was the lowest number of false positives when predicting genocide.

Appendix T: Rotated Principal Component Loadings For MI 2

The rotated principal component loadings shown below were generated using the 54 variables listed in Appendix K, with 178 exemplars taken from the second of the five multiply imputed data sets. The raw data was standardized, and subjected to principal component analysis via the "princomp" function in MATLAB. Only the first thirteen components were retained. The remaining 13 loadings were then rotated with a varimax rotation. The results are shown below.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
Year	-0.0012	-0.0027	-0.0784	-0.0277	-0.4654	-0.0732	-0.0333	-0.14	-0.0295	0.0451	0.0401	-0.0071	0.0123
Literacy	-0.0007	0.1916	0.0088	-0.0261	-0.2333	0.0015	0.2696	7.50E-03	0.0554	0.0706	0.0631	-0.0127	-0.0273
Gender Parity	0.0552	0.3808	0.0157	-0.0103	-0.1035	-0.1113	-0.0139	-0.0339	-0.0434	-0.0478	0.0667	-0.0607	0.0347
Primary Commodity Exports	-0.2143	-0.0387	-0.0686	-0.1081	0.2322	-0.0244	0.022	0.0652	0.1082	-0.0462	0.0568	-0.0277	0.0331
Forested Land	-0.0529	0.0644	0.3139	0.0162	-0.0719	0.0371	-0.1185	-0.0469	-0.0329	0.0485	-3.90E-03	-5.50E-03	0.0103
Life Expectancy	-0.16	0.1199	-0.008	-0.0697	0.0084	-0.2697	-0.0593	0.0538	0.0998	-0.1684	0.1609	-0.0181	-0.0096
Infant Mortality Rate	0.1198	-0.1594	-0.0306	0.0883	0.1545	0.3245	-0.0201	0.0336	-0.059	0.0212	-0.0391	-0.0014	0.0128
Youth Bulge	-0.2706	-0.0491	-0.1055	-0.1052	-0.0285	-0.091	-0.0301	0.122	0.1624	0.142	-0.0179	0.0132	0.0077
GDP per capita 98	0.2867	0.0424	0.0143	0.0511	-0.0989	-0.0481	0.0472	0.0511	-0.003	0.034	0.0632	-0.011	-0.0125
Trade Openness	0.1645	0.0199	-0.117	-0.0099	0.0451	-0.2724	-0.0367	0.0259	0.0596	-0.0255	-0.0081	-0.0146	0.03
Population	-0.0588	0.0337	0.0509	0.4042	-0.0545	0.0988	0.0329	-0.0273	0.0329	0.0233	0.0437	0.0131	-0.021
Urban	0.3071	-0.0865	-0.0338	-0.0646	-0.1241	0.0916	0.0655	0.0258	0.0522	0	0.0174	-0.0077	0.0052
Telephone Subscribers per 100	0.0752	-0.1137	-0.0583	-0.0489	-0.2226	-0.0737	0.1982	-0.0928	-0.0238	-0.0506	0.2351	0.0443	-0.0215
Foreign Aid as % GNI	-0.0888	0.062	-0.1641	-0.2419	-0.0932	0.106	-0.3826	0.1179	0.0191	0.0477	-0.0448	0.0033	0.0035
Military as % of GDP	-0.057	-0.0242	0.0191	-0.087	0.0084	-0.4913	-0.0182	-0.0442	-0.2336	0.0478	-0.1765	0.0654	0.0131
Agriculture as a % of GDP	-0.1396	0.1039	-0.0019	0.0516	0.024	0.2942	-0.1221	-0.0223	-0.12	-0.0402	0.0009	-0.0084	-0.0202
Durability	0.0092	-0.0107	-0.0911	-0.0063	-0.0953	0.0242	0.0643	0.5657	0.0384	0.0294	-0.0363	0.0602	-0.0458
Trade Ratio	0.0534	0.0949	-0.0249	-0.0096	0.0452	0.0021	0.5024	0.0359	-0.0381	-0.0206	-0.0567	-0.0125	0.0361
Foreign Aid per Cap	0.2385	-0.0184	-0.0983	-0.0547	0.0566	0.0128	-0.0816	0.0521	0.0569	0.0896	0.032	0.0553	-0.044
GDP Growth	-0.1067	0.0219	0.0489	-0.0202	-0.0858	-0.0257	-0.0661	-0.0184	0.0054	-0.3279	-0.0254	-0.3752	0.0158
Missing Data	-0.0334	-0.1648	0.0022	-0.1216	0.3755	-0.0478	0.1861	-0.1081	-0.0859	0.1044	-0.1074	-0.083	0.0846
Bad Neighbors	0.0491	0.2827	0.1928	0.0295	-0.0142	0.0524	0.0526	0.0323	0.0973	0.0668	-0.0561	0.1002	-0.0993
Ethnic Fractionalizations	0.1317	0.4409	-0.0079	-0.0364	-0.0084	0.216	-0.1086	-0.016	-0.0258	-0.0504	-0.0698	0.0072	0.052
Linguistic Fractionalization	-0.0823	0.3523	0.038	0.176	0.0652	-0.0613	0.0554	0.0243	-0.0472	0.0081	-0.0415	-0.0147	0.0035
Transition Government	-0.0116	-0.0417	0.003	0.0929	-0.0359	0.0057	-0.0701	-0.0361	0.0045	0.0352	-0.0517	-0.014	-0.8551
Anarchy	-0.033	0.0842	-0.0236	-0.2855	-0.0953	0.1593	0.2172	-0.3152	-0.256	0.0473	-0.2284	0.0117	0.0683
Full Autocracy	0.0435	0.0564	0.0637	-0.0624	0.1688	-0.0683	-0.0596	0.4713	-0.1361	-0.0403	0.0151	-0.0087	0.1324
Partial Autocracy	-0.0755	-0.1385	-0.0477	-0.0163	-0.0644	-0.0458	0.0349	-0.2732	0.0214	-0.0864	0.4721	0.059	0.072
Partial Democracy w/Factions	0.0445	-0.012	-0.0328	0.3155	-0.0877	0.0006	-0.0992	-0.1607	0.4381	0.085	-0.2573	-0.0547	0.326
Political Discrimination	0.055	0.0876	-0.0001	0.03	0.0489	0.1307	-0.038	0.1235	-0.1093	0.0925	0.5751	-0.0404	0.0312

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
Economic Discrimination	0.1758	0.0032	0.0948	-0.0635	-0.1244	-0.1617	-0.1096	0.0719	-0.0111	0.1643	0.0212	-0.0756	0.0665
Years since last conflict	0.1383	0.1441	-0.0792	-0.0508	0.3437	-0.0154	0.0286	-0.18	0.0922	0.0047	0.1532	-0.0012	-0.0577
Change in Calories	0.0284	0.0216	-0.0941	0.0024	0.0833	0.0236	0.0404	-0.0332	-0.0245	0.099	-0.0111	-0.8366	-0.0456
Change in IMR	-0.067	0.0422	-0.0667	-0.031	-0.0044	0.0048	0.0597	-0.0022	-0.0784	0.7118	-0.0773	0.0484	-0.0472
Pct Paved	0.0203	-0.0264	0.3572	0.0014	-0.0598	-0.0699	-0.0323	0.0168	-0.0397	0.0739	0.0303	-0.0258	-0.0038
Km Roads	-0.2085	0.0472	-0.1169	-0.0064	0.0198	0.0004	0.2001	0.0245	0.0747	0.0142	0.1195	0.0486	-0.0376
Calories	-0.0036	-0.0382	0.1448	-0.0371	-0.0292	0.0411	0.1631	0.1492	0.2998	0.0442	0.1511	-0.1442	-0.0148
Education as a % GNI	-0.0757	-0.0717	-0.2407	0.1581	0.1136	0.1142	0.1607	0.0454	0.1035	-0.0201	-0.0328	0.1027	0.0063
Water Per Capita	0.2734	0.0231	-0.0283	-0.0986	0.0596	0.002	-0.0762	-0.0446	0.0127	-0.0398	-0.0312	0.0132	0.0095
Population Density	-0.1029	0.17	-0.1898	0.2731	-0.0753	-0.1454	-0.0113	-0.0515	0.0022	-0.0018	-0.0053	0.0214	-0.0003
Arable Land Per cap	-0.0497	-0.0135	0.3644	0.0533	0.0677	0.0951	0.0931	0.0283	-0.0209	-0.0386	0.0107	0.012	-0.0009
Water/ AG interaction	0.285	0.0001	-0.0322	-0.047	0.0541	-0.0414	-0.0203	-0.0267	0.0129	-0.0201	-0.0276	0.0114	0.0055
Land Stress	-0.0575	-0.0176	0.3767	0.0113	0.0631	-0.002	0.1035	0.0165	0.0621	-0.0092	-0.0066	0.0236	1.60E-03
Water / Agriculture / Land	-0.0003	0.0194	0.3597	-0.098	0.0628	-0.0163	-0.0365	-0.0516	0.0719	-0.1242	-0.1029	0.0613	-0.0065
Road per Cap	-0.0511	-0.1014	-0.0394	-0.0422	-0.3047	0.0551	0.153	0.2323	-0.0074	-0.1242	-0.2002	-0.1533	0.0245
Relative GDP Per Cap	0.2821	0.0697	0.0385	0.0352	0.0428	-0.0766	0.0445	0.0078	-0.0245	-0.0016	0.0442	0.001	-0.0035
Somalia	-0.0532	0.0824	-0.0998	-0.252	-0.041	0.2984	-0.2111	-7.84E-02	0.0039	-0.0501	0.0138	0.0335	0.048
Kenya	-0.0884	0.4008	-0.1174	-0.0935	0.0977	-0.0211	0.147	0.0299	0.0756	0.0433	0.0197	0.0014	-0.0009
Ethiopia	0.0069	-0.0304	-0.05	0.4287	0.0131	0.1001	-0.0027	0.0066	-0.1431	-0.0495	-0.0466	-0.0022	0.0141
Djibouti	0.3135	-0.0606	-0.0657	0.0262	-0.0272	0.0293	0.0843	0.0427	0.0349	-0.0121	-0.0283	-0.0038	0.0031
4 Year Lagged Battle Deaths	-0.0607	-0.0269	-0.0133	0.1779	-0.0246	-0.0132	-0.0696	0.0621	-0.2814	0.072	0.0681	0.0636	0.2711
4 Year Lagged Refugees	-0.032	-0.0418	0.0369	0.0786	-0.0815	0.0619	0.0518	0.064	-0.5109	0.0284	0.068	-0.1002	0.0303
4 Year Lagged Genocide Deaths	-0.0443	-0.0422	0.1702	0.0199	-0.0195	-0.0104	-0.1569	-0.0449	0.1415	0.4033	0.1572	-0.1617	0.0776
4 Year Lagged Malnutrition	0.063	0.0511	-0.0865	0.21	0.1666	-0.2377	-0.1983	-0.0438	-0.1695	-0.0376	-0.0287	-0.0462	-0.0719

Appendix U: Error Logs for Refugees Per Capita

Note: Thresholds at the bottom of the table show refugees per million population. A -1 indicates the model predicted less than the threshold at the bottom when there were actually more, while a 1 indicates the model overestimated the actual number of refugees per million. 0's indicate a correct forecast. Refugees per million has been rounded to the nearest whole number to save space.

Yemen	2003	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Yemen	2004	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Yemen	2005	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Yemen	2006	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Somalia	2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Somalia	2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Somalia	2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Somalia	2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sudan	2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sudan	2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sudan	2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sudan	2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Djibouti	2003	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Djibouti	2004	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Djibouti	2005	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
Djibouti	2006	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
Kenya	2003	0	0	0	0	0	0	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kenya	2004	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kenya	2005	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kenya	2006	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0
Ethiopia	2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
Ethiopia	2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Ethiopia	2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	0	0	0	0
Ethiopia	2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	0	0
Eritrea	2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eritrea	2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eritrea	2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eritrea	2006	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Refugees per million		0	1	11	12	13	16	19	24	26	27	28	29	31	34	46	49	51	54	57	59	143	621	629	643	857	929	1429

Appendix V. Predictions For the Region 2007-2010

Note: For battle deaths per capita, refugees per capita, and genocide a 0 indicates the model forecasts less than the threshold discussed in Chapter 4 occurring in each country and year. The number listed beneath undernourishment is the predicted percentage of the population that will be defined as undernourished as defined by the UN FAO. Multiply imputed data set 2 was used to generate these predictions. The second table indicates the forecast number of refugees and battle deaths if the threshold is exceeded in the first table.

Country	Year	Battle Deaths Per Capita	Refugees Per Capita	Genocide	Malnutrition	Instability
Yemen	2007	0	0	0	22.5317	0
Yemen	2008	0	0	0	19.1	0
Yemen	2009	0	0	0	14.4285	0
Yemen	2010	0	0	0	9.1343	0
Somalia	2007	0	1	0	18.2328	1
Somalia	2008	0	1	0	9.7749	1
Somalia	2009	0	1	0	5.9229	1
Somalia	2010	0	1	0	19.1541	1
Sudan	2007	1	1	1	15.6679	1
Sudan	2008	1	1	1	10.6253	1
Sudan	2009	1	1	1	7.1761	1
Sudan	2010	1	1	1	3.8282	1
Djibouti	2007	0	0	0	29.891	0
Djibouti	2008	0	0	0	26.824	0
Djibouti	2009	0	0	0	17.0651	0
Djibouti	2010	0	0	0	17.2837	0
Kenya	2007	0	0	0	24.7594	0
Kenya	2008	0	0	0	18.8233	0
Kenya	2009	0	0	0	10.6466	0
Kenya	2010	0	0	0	5.9641	0
Ethiopia	2007	0	1	0	37.724	1
Ethiopia	2008	0	1	0	38.8035	1
Ethiopia	2009	0	1	0	32.5342	1
Ethiopia	2010	0	1	0	29.6439	1
Eritrea	2007	1	1	0	65.9558	1
Eritrea	2008	1	1	0	66.7145	1
Eritrea	2009	1	1	0	58.8768	1
Eritrea	2010	0	1	0	67.2961	1

Country	Year	Battle Deaths	Refugees
Yemen	2007	0	0
Yemen	2008	0	0
Yemen	2009	0	0
Yemen	2010	0	0
Somalia	2007	0	3611.336
Somalia	2008	0	3737.07
Somalia	2009	0	3866.233
Somalia	2010	0	3988.502
Sudan	2007	365.92664	16466.7
Sudan	2008	370.96866	16693.59
Sudan	2009	377.62842	16993.28
Sudan	2010	385.74264	17358.42
Djibouti	2007	0	0
Djibouti	2008	0	0
Djibouti	2009	0	0
Djibouti	2010	0	0
Kenya	2007	0	0
Kenya	2008	0	0
Kenya	2009	0	0
Kenya	2010	0	0
Ethiopia	2007	0	31332.64
Ethiopia	2008	0	32101.46
Ethiopia	2009	0	32873.98
Ethiopia	2010	0	33650.09
Eritrea	2007	44.70145	2011.565
Eritrea	2008	45.54408	2049.484
Eritrea	2009	46.69638	2101.337
Eritrea	2010	0	2154.147

Thresholds exceeded in positive predictions, converted to raw numbers

Appendix W: Discussion of Development of a Instability Index

The dependent variables in this study; battle deaths per capita, genocide and politicide deaths per capita, refugees per capita, and undernourishment as a percent of the population were used as dependent variables. In order to use regression on the variables, a single index of the instability indicators needed to be created, and the output had to be approximately normally distributed. There is limited correlation between the dependent variables. Figure 3.10 shows the correlation matrix describing the relationship between the dependent variables and the t-score. Note that the variables were transformed to have reduce the difference in weighting between them. Some disparity still remains in correlation between the original dependent variables and the t-score. In addition, note that the t-score does not have a linear relationship with a combination of the other dependent variables.

	Battle.Deaths	Refugees.	Genocide	Malnutrition	T-score
Battle.Deaths	1	0.3908	0.3793	0.0173	0.7
Refugees.	0.3908	1	0.1644	0.2077	0.702
Genocide	0.3793	0.1644	1	-0.2014	0.503
Malnutrition	0.0173	0.2077	-0.2014	1	0.451
T-score	0.7001	0.7018	0.5033	0.4508	1

Table W-1. Correlations Between Dependent Variables

The variables were transformed to approximate normal distributions to satisfy both Factor Analysis and OLS normality assumptions. All of the continuous data was normalized where possible. After transformation each data point was converted to a z score where:

$$z = \frac{y - \mu}{\sigma}$$

where

z is the distance from the mean of a normal distribution expressed in units of standard deviations

y is the normalized data point

μ is the mean of the normalized observed variable

σ is the standard deviation of the variable represented by y

These scores were summed for each country and year, to create a data set of 206 index data points. Their normality was checked using the BestFit software package, and the results are shown in Figure 3.11. Appendix G shows sample results from the index. It should be noted that the raw MI datasets did not require the data to be normalized.

Appendix X: Notes on Normalizing Distributions

In order to determine if normality assumptions for each variable are met at some level of confidence, Goodness Of Fit (GOF) tests were used to judge the transformations tested in the best fit program against one another. Wherever necessary, the attempt was made to ensure the variables in the study's dataset met the normality condition, or were at least transformed to better fit the assumption. There are limits on what can be done to transform a variable towards normality, however; it is noted where normality could not be truly obtained.

Three types of GOF tests were used: Chi-Squares (χ^2), Anderson-Darling (A-D), and Kolmogorov-Smirnov (K-S). Each of these tests the hypothesis:

H_0: Fail to reject the assumption the data is drawn from a normal distribution

H_1: Reject the assumption the data is drawn from a normal distribution

The Chi-Squared test is the oldest GOF test (Law, 2007: 341). It is not used as often as the others due to its tendency to frequently reject when there are a large number of samples, and usually accept the null hypothesis when there is a small sample. The test starts by separating data points into k bins, and then compares how many are in each bin with how many would be expected to be in those bins given a particular distribution. The Chi-Squares test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

where

n is the total number of observations

n_i is the number of data points in bin *i*

p_i is the probability an observation will fall into bin *i* given a particular distribution

The test statistic is a random variable with a Chi-Squared distribution with k-1 degrees of freedom. As the test statistic increases, it becomes more likely that the null hypothesis is rejected (Wackerly, *et al*, 2002: 684).

The K-S GOF is more commonly used now than the Chi-Squared. The K-S finds the difference between the empirical distribution function with the hypothesized distribution. The K-S test statistic is defined as:

$$D_n^+ = \max_{1 \le i \le n} \left\{ \frac{i}{n} - \hat{F}\left(X_i\right) \right\} \qquad D_n^- = \max_{1 \le i \le n} \left\{ \hat{F}\left(X_i\right) - \frac{i-1}{n} \right\}$$

$$D_n = \max\left\{D_n^+, D_n^-\right\}$$

Where

$\hat{F}$ is the cdf of the hypothesized distribution

X_i is the value of the i^{th} observation

n is the total number of observations

This equation simply expresses the maximum vertical distance between $F_n(x)$ and $\hat{F}(x)$. The test looks for some constant $d_{n,1-\alpha}$ where α is the level of the test. The null is rejected if some D_n exceeds $d_{n,1-\alpha}$ (Law, 2007: 347-349).

The A-D GOF test is similar to the Chi-Squared test in that it is able to test any given distribution where a cumulative distribution can be found. It is better than the (K-S) test at testing for differences in the tails of distributions (Banks, *et al*, 2005: 333). The A-D statistic is defined by:

$$A^2 = \frac{-(\sum_{i=1}^{n}(2i-1)(\ln Z_i + \ln(1 - Z_{n+1-i}))}{n} - n$$

where

n is the total number of observations

$\hat{F}$ is the cdf of the hypothesized distribution

X_i is the value of the i^{th} observation

$Z_i = \hat{F}(X_i)$ for I = 1,2,…,n

Again, the larger the value of A^2 the greater the chances the null hypothesis will be rejected (Law, 2007, 351-352).

Each of the tests generates a p-value. This value represents the probability that if a sample was drawn from the specified hypothetical distribution the sample would have a test value greater than or equal to the test statistic. Thus, a very small p-value coming from the three tests discussed indicates it is unlikely the observed distribution is drawn from the hypothesized one, in this case the normal.

Given the data set used in this study transformed data was used when it provided the smallest possible p-value amongst the tests described above using the BestFit software. Not all data could be normalized. For example, the data for kilometers of roads appears bimodal. The transformations used, and the normality of the data used is shown in Appendix I.

The end state goal of creating an index as a dependent variable to represent failing state indicators on a continuous scale was for the index to have an approximately normal distribution to regress against, and secondly, it needed each of the variables to be approximately equally weighted on the index. When the data was converted to z-scores, it resulted in an index which possessed neither property. Initially, the difference between the highest and lowest z-score for each variable was very different resulting in unequal weightings of the stability indicators.

	Battle Deaths Per Capita	Refugees per Capita	Genocide Per Capita	Undernourishment
Minimim z	-0.2568	-0.5802	-0.2722	-1.8575
Maximum Z	10.6825	4.9938	9.5921	2.6462
Difference	10.93937	5.57409	9.8644	4.5037

Table X-1. Initial Difference In z-scores

To reduce the difference between the variables, the raw data was transformed with exponents. A pilot study was used to find exponents which resulted in similar differences between variable z-scores. The exponents do not represent any interpretable number, the fit of the model was the primary goal. As a result, the variables were raised to the powers in table X-2, resulting in much more evenly weighted factors in the overall score.

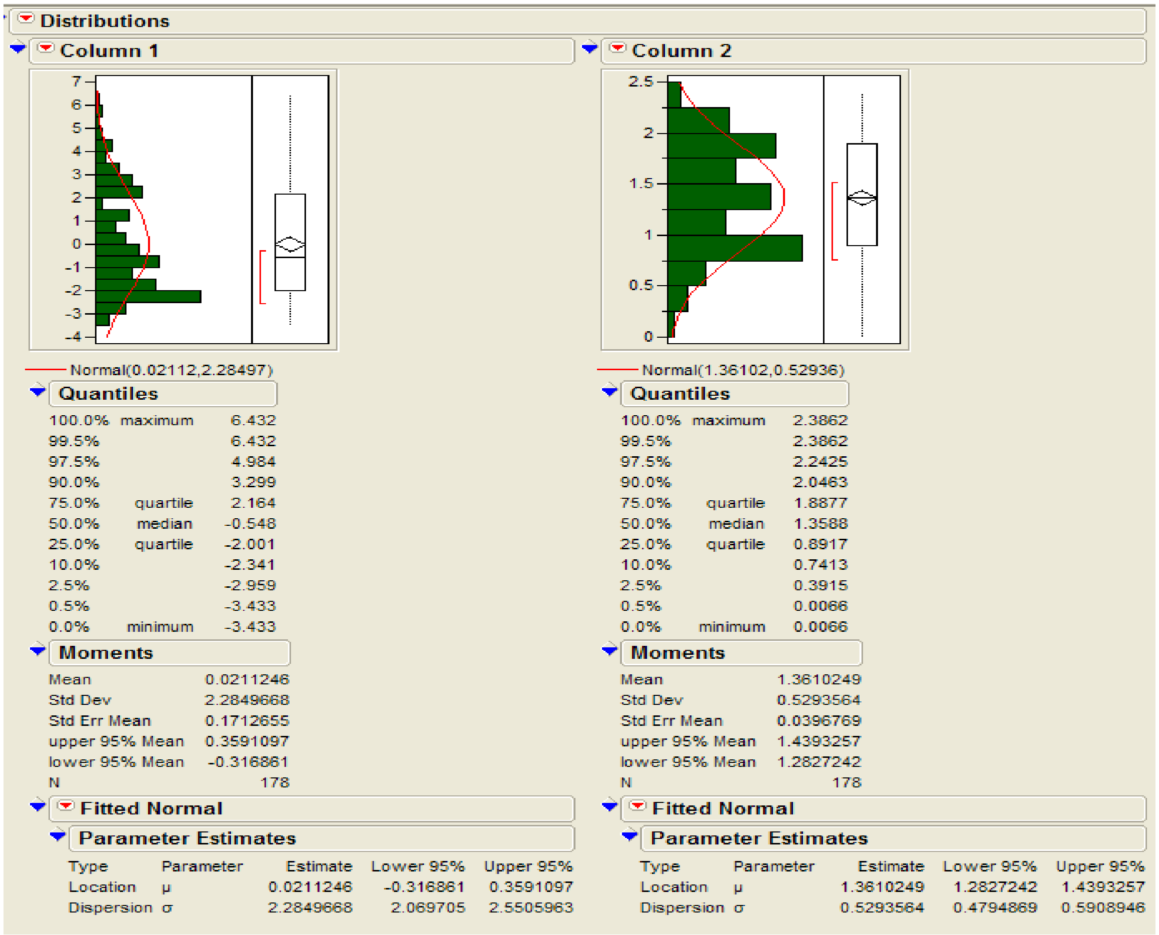
	Battle Deaths Per Capita	Refugees per Capita	Genocide Per Capita	Undernourishment
Power	^.2	^.75	^.25	^.25
Minimim z	-0.8928	-0.6752	-0.4013	-2.8188
Maximum Z	4.1471	4.3979	4.5193	2.1445
Difference	5.0400	5.0731	4.9206	4.9634

Table X-2. Final Differences In z-scores

The second issue was the normalization of the resultant t-score. It was highly skewed, and displays a bimodal distribution. To remove the skew, the data was transformed with the following:

Transformed t-score = ln ((Raw score) + (minimum raw score + 1))

This both base-lined the index at zero and removed a significant portion of the skew. The bimodality of the data still exists; however, initial tests of the data indicated the non-normal properties would not prevent the creation of a useful and significant model. The graphic in Figure 3.13shows the distribution of the t-score data before and after transformation.



Distributions

Column 1

Normal(0.02112,2.28497)

Quantiles

100.0%	maximum	6.432
99.5%		6.432
97.5%		4.984
90.0%		3.299
75.0%	quartile	2.164
50.0%	median	-0.548
25.0%	quartile	-2.001
10.0%		-2.341
2.5%		-2.959
0.5%		-3.433
0.0%	minimum	-3.433

Moments

Mean	0.0211246
Std Dev	2.2849668
Std Err Mean	0.1712655
upper 95% Mean	0.3591097
lower 95% Mean	-0.316861
N	178

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	0.0211246	-0.316861	0.3591097
Dispersion	σ	2.2849668	2.069705	2.5505963

Column 2

Normal(1.36102,0.52936)

Quantiles

100.0%	maximum	2.3862
99.5%		2.3862
97.5%		2.2425
90.0%		2.0463
75.0%	quartile	1.8877
50.0%	median	1.3588
25.0%	quartile	0.8917
10.0%		0.7413
2.5%		0.3915
0.5%		0.0066
0.0%	minimum	0.0066

Moments

Mean	1.3610249
Std Dev	0.5293564
Std Err Mean	0.0396769
upper 95% Mean	1.4393257
lower 95% Mean	1.2827242
N	178

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	1.3610249	1.2827242	1.4393257
Dispersion	σ	0.5293564	0.4794869	0.5908946

Figure X-1. Instability Index Distribution Before and After Transformation (JMP)

Appendix Y: A Description of ARIMA (*p,d,q*) Models

This sub-section uses the Makridakis, Wheelwright, and McGee test to show the development of ARIMA models. (Makridakis, Wheelwright, and McGee, 1983). The basic ARIMA (*p, d, q*) model is best understood at a high level by describing *p, d*, and *q*. The autoregressive term *p* describes how many prior terms should be used to find the predicted value of Y_t. In practice, a ARIMA (1,0,0) (or AR(1)) model follows the form:

$$Y_t = \varphi_1 Y_{t-1} + \mu' + e_t$$

$$\mu' = (\mu - \varphi_1 \mu)$$

where

ϕ_1 is the autoregressive coefficient for a one time unit lag with a value between -1 and 1.

μ is the mean of all responses.

e_t is error at a particular iteration.

For AR(*p*) models the general form becomes:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-1} + ... + \varphi_p Y_{t-p} + \mu' + e_t$$

$$\mu' = (\mu - \varphi_1 \mu - \varphi_2 \mu - ... - \varphi_p \mu)$$

(Makridakis, Wheelwright, and McGee, 1983, 359-364)

The values of φ are chosen to minimize the Mean Square Error (MSE) of the model and are calculated by the JMP program. Since the value Y_t is found using coefficients on other variables in the time series plus a mean and an error term, the model resembles an OLS model with *p* independent variables, hence the description of this model as Auto(correlation) Regressive. The exact values of φ can be found via linear programming model when making an assumption regarding the normality and looking to minimize MSE. The optimal value of φ minimizes the sum of squares residual when comparing the model with the actual time series data. The formulation and proof of this minimization algorithm in JMP is detailed in Box and Jenkins, 1976 Appendix 7.5 (Box and Jenkins, 1976, 243-284).

The differencing term *d*'s purpose in an ARIMA model is to account for a non-stationary process, meaning that the mean of the function changes throughout the time series data. Figure 3.3 and 3.4 shows examples of stationary and non-stationary processes. An ARIMA (0,1,0), or I(1), model is formulated by:

$$Y_t = Y_{t-1} + e_t$$

This model attempts to estimate the error term by observing at the previous error and compensating for it by adding the expected error to the previous term. For example, given notional time series data points 2, 4, 6, 8,…, 20, the estimate of each error term would be 4-2, 6-4, etc leading to a series of error estimates 0, 2, 2,…., 2. When forecasting this series the estimates would be $Y_t = Y_{t-1} + e_t$ where e_t is the expected error, $e_t = Y_{t-1} - \hat{Y}_{t-1}$. In the example of the notional data, the estimates would be 2, 2, 4,…, 18. Of note is the one time period lag (Makridakis, Wheelwright, and McGee, 1983, 359).

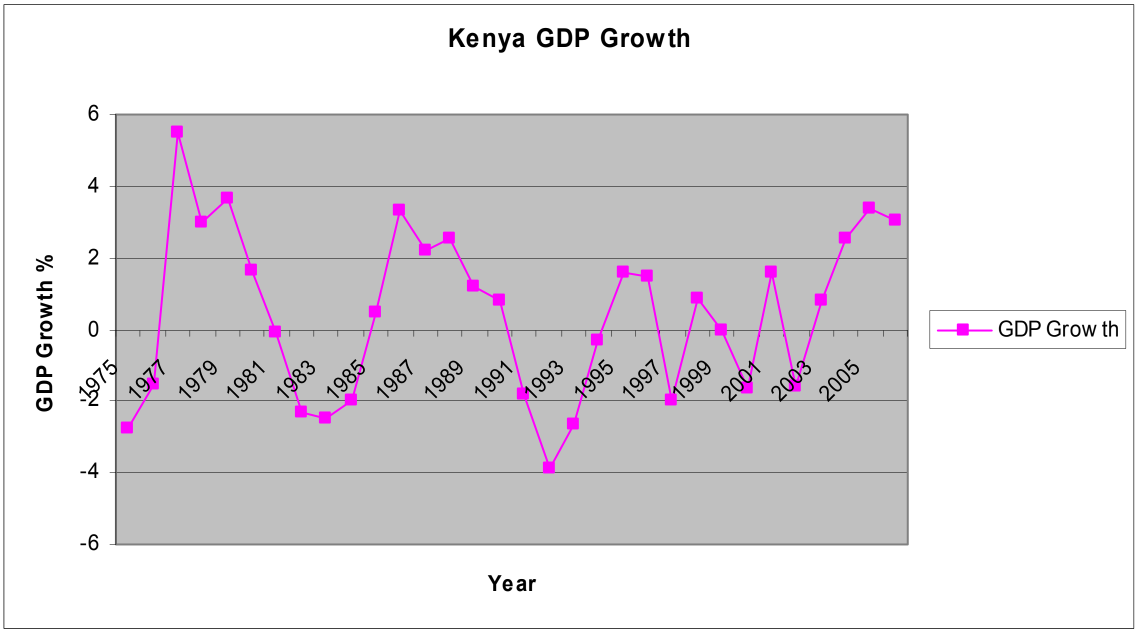


Figure Y-1: Example of a Stationary Process

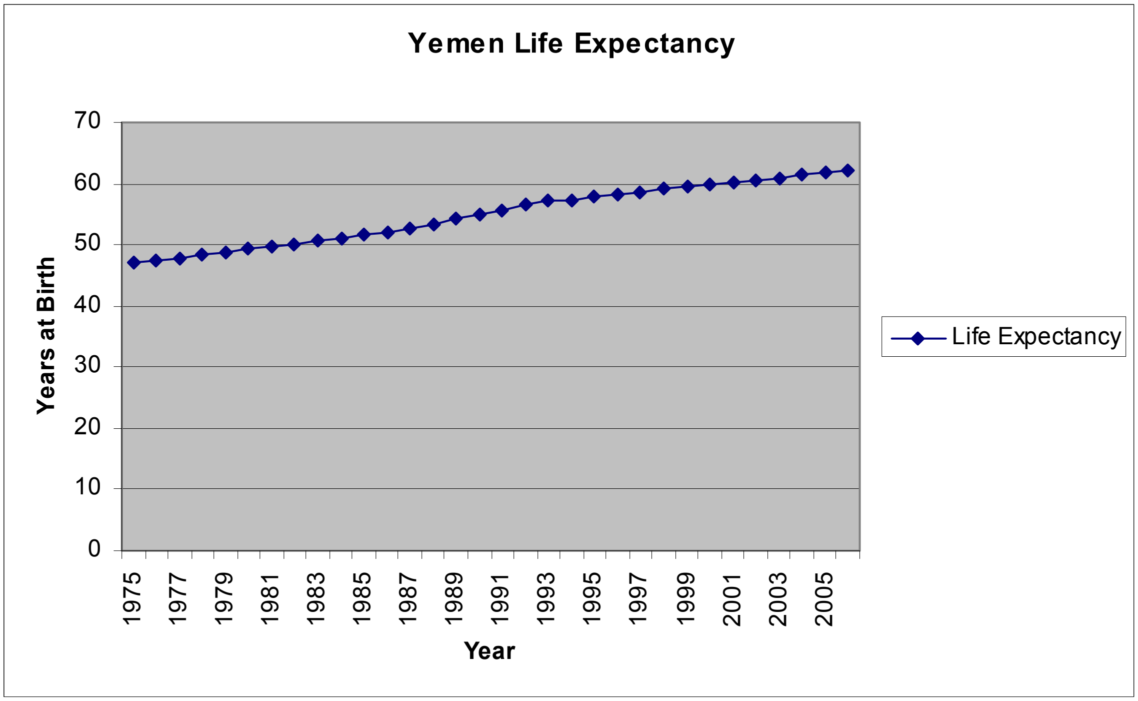


Figure Y-2. Example of a Non-Stationary Process

Another way of expressing this concept is the model finds the difference between the last two observations, and tries to compensate for the trend. Makridakis shows $X_t^{'} = X_t - X_{t-1}$. This series has n-l values, and will be stationary if the trend in the original data X_t is linear. If the values of $X_t^{'}$ are autocorrelated, then the degree of differencing should be increased to 2. This is found by $X_t^{''} = X_t^{'} - X_{t-1}^{'}$, resulting in n-2 observations. This can be continued until the autocorrelations approach zero after the second or third lag period (Makridakis, Wheelwright, and McGee, 1983, 380-383). However, in practice it is rare to need to go beyond d =2. In this study, no model of the data benefitted from d > 2, which indicates the mean of the data over time can always be expressed as a quadratic function (or less), and does not require higher degree polynmials to be modeled. Figure 3.5 shows a case from the data set where the rate of mean change was not linear.

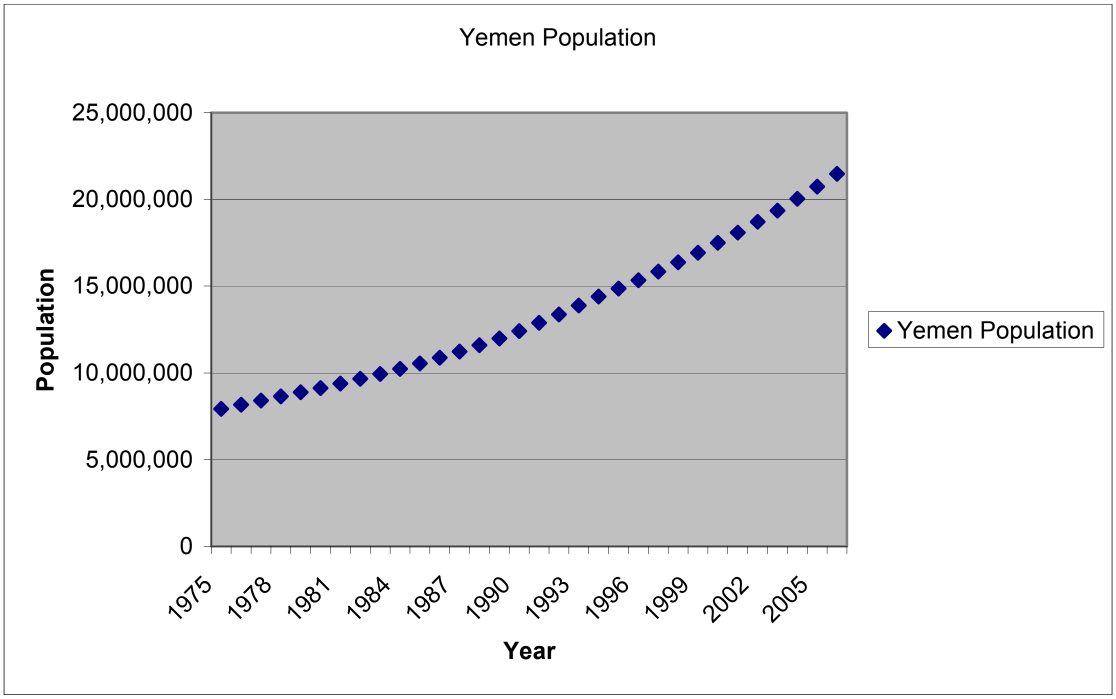


Figure Y-3. Example of Non-Stationary Process with *d* > 1

Differencing is another way to compensate for autocorrelation, but the difference between an ARIMA (1,0,0) model and an ARIMA (0,1,0) model is the former will deviate back towards the mean by assuming it is stationary, and the latter assumes a non-stationary mean. An integrated ARIMA (1,1,0) model follows the form:

$$\hat{Y}_t = \mu + Y_{t-1} + \varphi_1 (Y_{t-1} - Y_{t-2})$$

This model keeps μ as a constant, but it also deviates from it based on a differencing term and the autocorrelation between previous observations (Nau, 2005, 2).

The final term in the ARIMA (*p,d,q*) is *q*, the number of lagged forecast errors in the prediction. The *q* term defines how many previous error terms are used to predict the next value of X_t. The simple Moving Average (MA) (1) model, or ARIMA (0,0,1) is expressed as:

$$Y_t = \mu + e_t - \theta_1 e_{t-1}$$

Where

θ_1 is the moving average coefficient for a one time unit lag with a value between -1 and 1.

μ is the mean of all responses.

e_t is error at a particular iteration (Makridakis, *et al.* 1983, 380-383). A generalized form of the MA(q) model can be written as:

$$X_t = \mu + (1 - \theta_1 X_{t-1} - \theta_2 X_{t-2} - ... - \theta_q X_{t-q})e_t$$

Where

$(1 - \theta_1 X_{t-1} - \theta_2 X_{t-2} - ... - \theta_q X_{t-q})e_t$ represents the weighted moving average difference from the mean (Makridakis, *et al.* 1983, 421).

The value range of φ_p and θ_q depends on the value of p and q. In the simple ARMA (1,1) model φ_1 and θ_1 are between -1 and 1. However, in AR(2) and MA(2) models φ_1 and θ_1 are between -2 and 2, and φ_2 and θ_2 are between -1 and 1. This pattern continues for larger values of p and q (Makridakis, *et al.* 1983, 383).

From these basic parameters p, d, and q there are a large number of combinations possible. In practice, though, it is unusual to exceed 3 for p and q, and 2 for d. The most basic combination of all three of these is an ARIMA (1,1,1) model:

$$(1 - B)(1 - \varphi_1 B)X_t = \mu' + (1 - \theta_1)e_t$$

$$\mu' = \mu - \phi_1 \mu$$

Where

$B = X_{t-1}$

$(1 - B)$ represents the first difference for $d = 1$

$(1 - \varphi_1 B)$ represents the AR (1) portion of the model

$(1 - \theta_1 B)$ represents the MA (1) portion of the model

This reduces to:

$$X_t = (1 + \varphi_t)X_{t-1} - \varphi_1 X_{t-1} + \mu' + e_t - \theta_1 e_{t-1}$$

This resembles a regression equation, except for the multiple error terms on the right hand side. These error terms are responsible for the autoregressive portion of the model (Makridakis, *et al.* 1983, 426). More complicated mixed models are built along this same pattern. For example, the starting point for deriving the model of a ARIMA (2,2,2) model is:

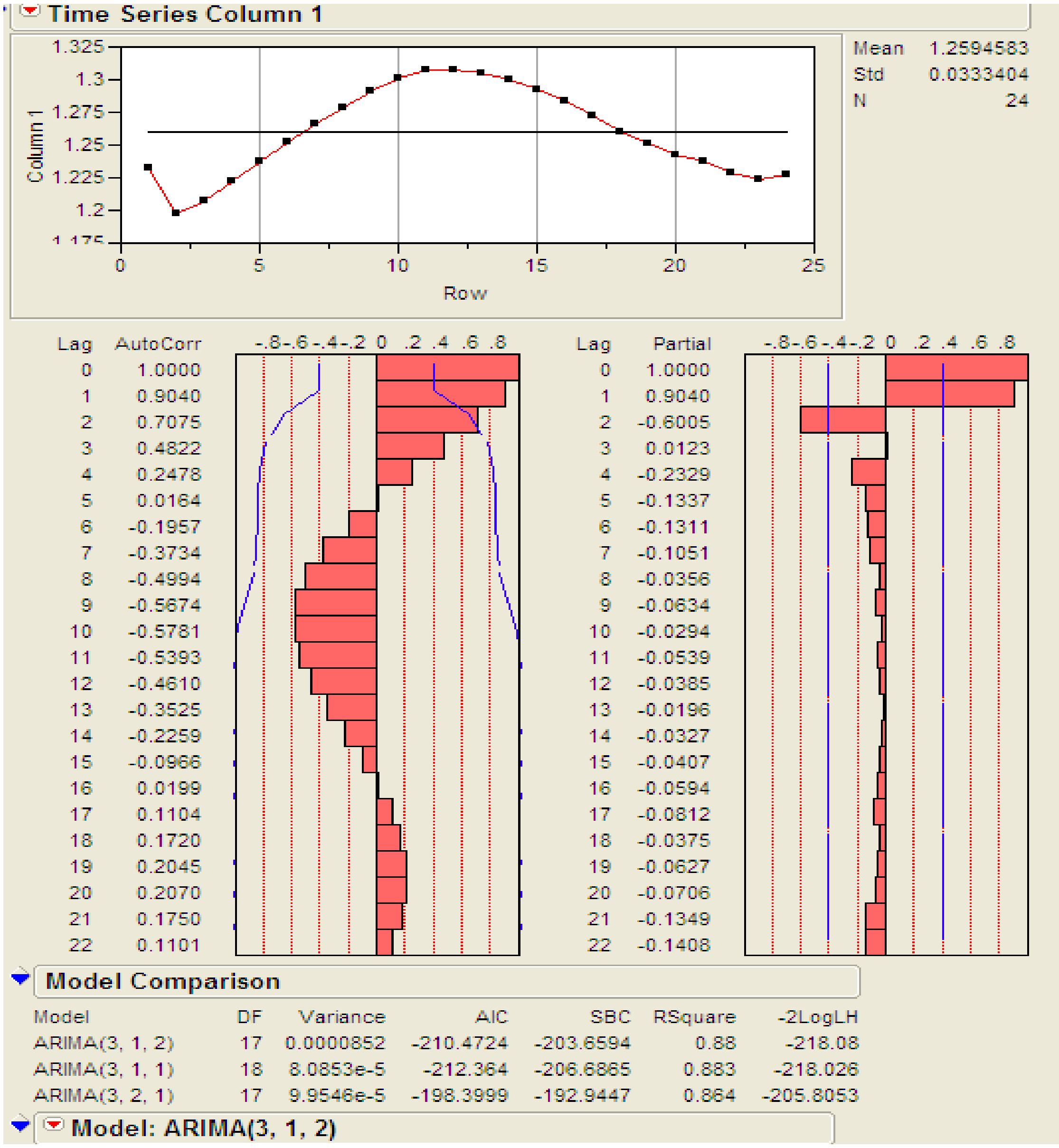
$$(1 - \varphi_1 B - \varphi_2 B^2)(1 - B)^2 X_t = (1 - \theta_1 B - \theta_2 B^2)e_t$$

Where $B^n = X_{t-n}$ (Makridakis, *et al.* 1983, 480).

This study applied ARIMA models to data which came within six years at either end of the studied time period. Certain self restrictions were placed on the use of extrapolation:

1. R_{adj}^2 >= .5. If a model of this fidelity could not be generated, then the missing data imputation was left to the "catch all" of multiple imputation, which is discussed in Chapter 3.2.7. The selection of this number is simply based on the idea that the model explains the majority of the variance while maintaining some level of parsimony.
2. No impossible values generated (i.e. there cannot be negative literacy rates). There is justification for deliberately developing and using values which are known to be incorrect, or have no chance of being correct. In some cases where ARIMA was having difficulty with tails that dipped below zero, the data set was truncated to look only at the tails themselves. These instances are noted in Appendix E.
3. A limit of no more than 6 years of data generated going forwards or backwards from the last data point. For instance, given a notional example of Yemen if it had data points for literacy from 1980-1997, then extrapolation was used to determine the values for 1975-1979, and the data for 1998-2006 was left for multiple imputation because it exceeded a six year prediction. Pilot analysis indicated going beyond this typically made the 95% confidence interval unacceptably large.

Each series of data potentially eligible to be extrapolated was examined in JMP. Figure 3.6 shows the basic output of a time series analysis in JMP. The top graph provided a general indication of whether the mean was stationary. The autocorrelation and the blue line indicating what lags were significant according to the Durbin-Watson test, and suggests what levels of p should be tested. The partial correlation graph on the right suggested what level of q should be selected. In each case, multiple models were tested against each other to determine if any met the minimum R_{adj}^2 criteria, and if so, which model provided the largest R_{adj}^2. Given similar R_{adj}^2 the model with the lower variance was selected. R_{pred}^2 was not provided by the program, but could be calculated, and is described in Chapter 3.5. Give the time consuming nature of individually extrapolating the data set, R_{adj}^2 was judged to be sufficient for determining model efficiency. Following extrapolation, 11.6% of the data was still missing, and would have to be developed via multiple imputation.



Mean	1.2594583
Std	0.0333404
N	24

Lag	AutoCorr
0	1.0000
1	0.9040
2	0.7075
3	0.4822
4	0.2478
5	0.0164
6	-0.1957
7	-0.3734
8	-0.4994
9	-0.5674
10	-0.5781
11	-0.5393
12	-0.4610
13	-0.3525
14	-0.2259
15	-0.0966
16	0.0199
17	0.1104
18	0.1720
19	0.2045
20	0.2070
21	0.1750
22	0.1101

Lag	Partial
0	1.0000
1	0.9040
2	-0.6005
3	0.0123
4	-0.2329
5	-0.1337
6	-0.1311
7	-0.1051
8	-0.0356
9	-0.0634
10	-0.0294
11	-0.0539
12	-0.0385
13	-0.0196
14	-0.0327
15	-0.0407
16	-0.0594
17	-0.0812
18	-0.0375
19	-0.0627
20	-0.0706
21	-0.1349
22	-0.1408

Model	DF	Variance	AIC	SBC	RSquare	-2LogLH
ARIMA(3, 1, 2)	17	0.0000852	-210.4724	-203.6594	0.88	-218.08
ARIMA(3, 1, 1)	18	8.0853e-5	-212.364	-206.6865	0.883	-218.026
ARIMA(3, 2, 1)	17	9.9546e-5	-198.3999	-192.9447	0.864	-205.8053

Figure Y-3. Example of JMP Time Series Analysis Output

The mean (1.259) , standard deviation (.0333), and number of observations (24) is shown in the upper right hand corner. The bar graph on the left indicates a lag on the AR component of the model of either 2 or 3 should be used, while the bar graph on the right suggests testing models using *q* values of 1 or 2. The graph at the top of the data over time suggests that the mean may be moving over time, albeit slowly, This suggest a *d* value of either 0 or 1. The model comparison field at the bottom shows the fit parameters of three models tested: ARIMA (3,1,2), (3,1,1), and (3,2,1). The ARIMA (3,1,1) model was selected, since it had the highest Rsquare, and the lowest variance.

Appendix Z: Continuous Models of Each of the Instability Indicators

Given a complete multiply imputed data set, canonical correlation and PCA scores were used to try to create 4 year forecasts of battle deaths, refugees, genocide deaths, and malnutrition. Both per capita and total (raw) numbers were used to test the former. Only the results for the raw numbers are shown in the tables and appendices, since the continuous models proved less useful than the discrete forecasting ones, and per capita models did not perform any better than the continuous models of continuous raw data.

Earlier attempts at forecasting using an index based on a combination of dependent variables were abandoned for several reasons. Firstly, in order to be of greater use to the sponsors of this research an index with no intrinsic meaning would only serve to add an unnecessary level of complexity to interpreting the results of a prediction. It also could not be independently verified and validated based on subject matter expert input in time. The proposed t-score was also poorly correlated with existing scores that could be found, particularly the Fund For Peace scores. Thus, continuous models attempted here modeled each individual instability indicator.

Z.1 Canonical Correlation Scores

Multiply Imputed Data Set 1 was used for initial trials, and a modified version of the MATLAB "canoncorr" function was used to generate the canonical correlation scores. The MATLAB function was modified to allow the creation of new scores using hold out data for the purpose of evaluating the method's predictive value. Two variables were removed from the training data since they caused the loadings matrix to be singular: Partial Democracy without Factionalism and Religious Fractionalization. Appendix K shows the variables used to generate the training independent data matrix. Training data for each country from 1975 until 1998 was placed in one group of variables, and malnutrition, battle deaths, refugees, and genocide data from 1979 through 2002 was placed in the other. Country data from 1999-2002 and instability indicator data from 2003-2006 was held out of the model to test the predictive ability of the model. The result was a 150 x 4 canonical correlation scores matrix which was used to create a regression model for each stability indicator. Part of these scores can be found in Appendix L.

When the canonical correlation scores of the training data was fitted via OLS to each individual dependent stability variable it produced models with the following results:

	Battle Deaths	Refugees	Genocide Deaths	Malnutrition
Rsquare	0.5546	0.8979	0.7168	0.9611
Rsquare Adj	0.5423	0.8951	0.709	0.9601
Rsquare Pred	0.4957	0.8847	0.6643	0.9585
RMSE	3839.1	1.42E+05	25752	3.0925

Table Z-1. Canonical Correlation Training Data Fit

In order to test each of these models, holdout data from 1999-2002 was used to try to predict battle deaths, refugees, genocide deaths, and malnutrition for each country during the years 2003-2006. The hold out scores were found by multiplying the hold out data times the 54 x 4 transformation matrix A, and adding preceding row of 1's added to the first column in the set of new scores. The new scores matrix was multiplied by the regression coefficients vector $\hat{b}$ to generate a set of predictions for each of the dependent indicator variables. Tablee Z-2 shows the results.

		Battle Deaths		Refugees		Genocide Deaths		Malnutrition	
Country	Year	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Yemen	2003	0	448.6811	550	-55657.7	0	15416.501	37	34.3166
Yemen	2004	0	2854.98	526	89007.74	0	29456.546	38	38.1005
Yemen	2005	0	897.7045	395	-104990	0	94446.867	30.4838	36.092
Yemen	2006	0	2501.998	573	-292948	0	207846.79	36.6368	34.3199
Somalia	2003	0	212.0109	315114	326100.6	0	-24639.74	16.1	23.5156
Somalia	2004	0	-1736.41	306200	128670.3	0	25833.799	19.7	19.6933
Somalia	2005	0	-3153.88	314066	33018.29	0	38607.542	12.7	19.0525
Somalia	2006	547	-294.824	388046	-15638.7	0	40884.176	18.6	16.1868
Sudan	2003	3225	4181.717	580727	486276	96000	41809.583	27	29.0543
Sudan	2004	6569	6058.429	696657	458697.8	192000	57219.398	26	31.8124
Sudan	2005	1204	4157.114	655732	267331	48000	76037.883	29.2417	30.5496
Sudan	2006	1002	5332.481	645093	191826.7	48000	136224.32	32.3014	28.9256
Djibouti	2003	0	8645.112	78	165302.6	0	-61737.48	26	24.7069
Djibouti	2004	0	9131.109	73	161010.7	0	-54278.67	24	22.5199
Djibouti	2005	0	11579.77	77	303619.9	0	-57258.95	25.2437	18.2103
Djibouti	2006	0	10982.71	105	364388.5	0	-18427.15	27.5347	15.1488
Kenya	2003	100	1519.429	883	50330.85	0	-12762.41	31	33.4802
Kenya	2004	52	-100.509	935	-84547.1	0	7788.3616	31	29.5709
Kenya	2005	251	-1485.88	1186	-235592	0	113843.1	28.2095	30.6891
Kenya	2006	567	261552.5	1753	-9.6E+07	0	-10689051	27.6009	33.7472
Ethiopia	2003	970	-3501.07	43676	-355462	0	-17750.32	46	40.94
Ethiopia	2004	936	2806.777	44219	-307389	0	-7354.137	46	42.3055
Ethiopia	2005	773	2927.382	46824	-298515	0	-1789.684	48.6663	44.5684
Ethiopia	2006	0	870.7391	65361	-431079	0	13697.817	48.366	44.1047
Eritrea	2003	57	-1580.38	116964	211976.9	0	25860.283	73	64.3102
Eritrea	2004	0	2459.05	121937	225000	0	41399.592	75	68.7003
Eritrea	2005	0	10064.3	130544	349482	0	-14645.17	72.2599	71.0569
Eritrea	2006	0	10672.17	168682	382230.7	0	2263.9724	71.7162	71.7176
Rsquared		0.485199348		0.491286812		0.500609111		0.915213343	
RMSE		51456.4474		18878158.18		62733.78037		4.984682214	

Table Z-2. Instability Indicator Predictions Using Canonical Correlation and OLS

The results shown above are problematic. First, the models for battle deaths, refugees, and genocide generate negative forecasts which are difficult to interpret. The other serious problem is the RMSE. The variance of the hold out data predictions is large enough that little insight is given to the forecaster.

There was a suspicion the non-constant variance of the battle death, refugee, and genocide models, along with non-normal dependent data, was behind these very high variances. Figure Z-1 shows the non constant variance in the battle deaths model. Refugees and genocides follow a similar pattern.

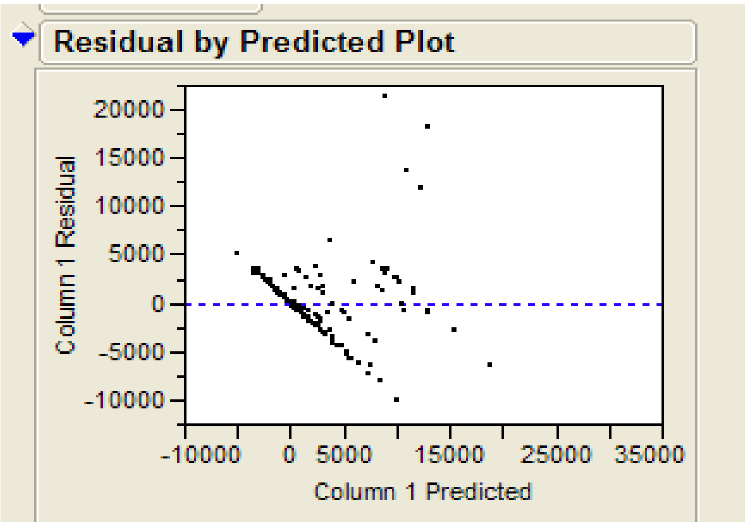


Figure Z-1. Residual Plot of Battle Death Predictions Using JMP

In order to compensate for the non-normality of the dependent data, as well as the no-constant variance, generalized linear models and generalized least squares were both attempted. Appendix M shows the results of several models that were tried on the battle deaths hold out data. Note that the data in Appendix M shows some of the other more complicated regression models have a lower RMSE, however the magnitude of these RMSE's is still large enough that the forecasting models of battle deaths, refugees, and genocide are of limited utility. Even in the best case scenario, a 95% prediction interval using Kriging surrogates with second order polynomials and an exponential correlation model would be +/- 24622.1 battle deaths. Of particular note is the predictions generated by OLS regression of the canonical correlation scores is nearly identical to a simple OLS regressions of the initial raw data. The results for forecasting models of refugees and genocide deaths were similarly unpromising, and are shown in Appendix Q. This appendix shows the results using canonical correlation scores, PCA scores, and raw data using OLS and the best DACE model in terms of RMSE.

As a final attempt to determine if canonical correlation might hold some useful forecasting power, the results were tested to see if the model could correctly categorize data by scoring the predictions the same way a logistic regression or DA prediction would. If the model prediction was below the threshold value, then it was scored a 0, and 1 if above the threshold. The categorizations were compared with the actual results, prediction errors were recorded and tallied, and then the process repeated for another (higher) threshold value. For almost all threshold values over the instability indicators of interest use of canonical correlation scores in using the best case RMSE model from DACE produced worse categorization results than DA and logistic regression. Figures Z-2 through Z-4 show the results using data from MI1. The only model showing promise was using canonical correlation as a genocide discriminator. However, using DA on PCA scores eventually proved superior, and these discrete models are discussed in section 4.3.

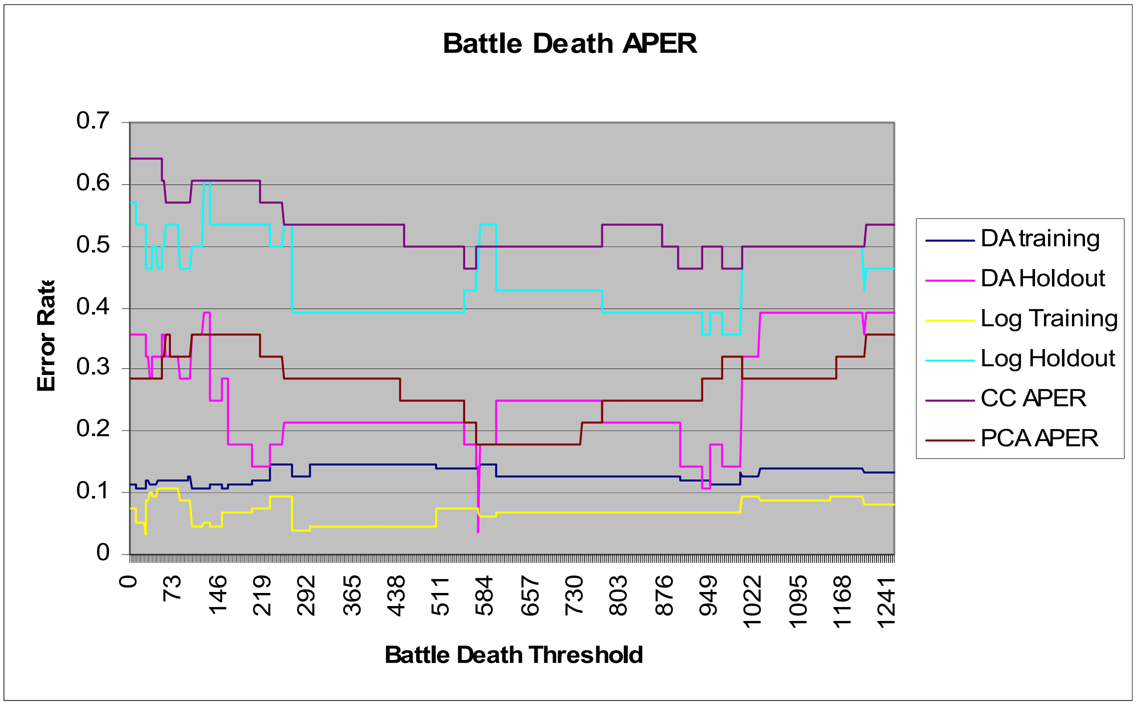


Figure Z-2. Battle Death Apparent Error Rates for Several Models

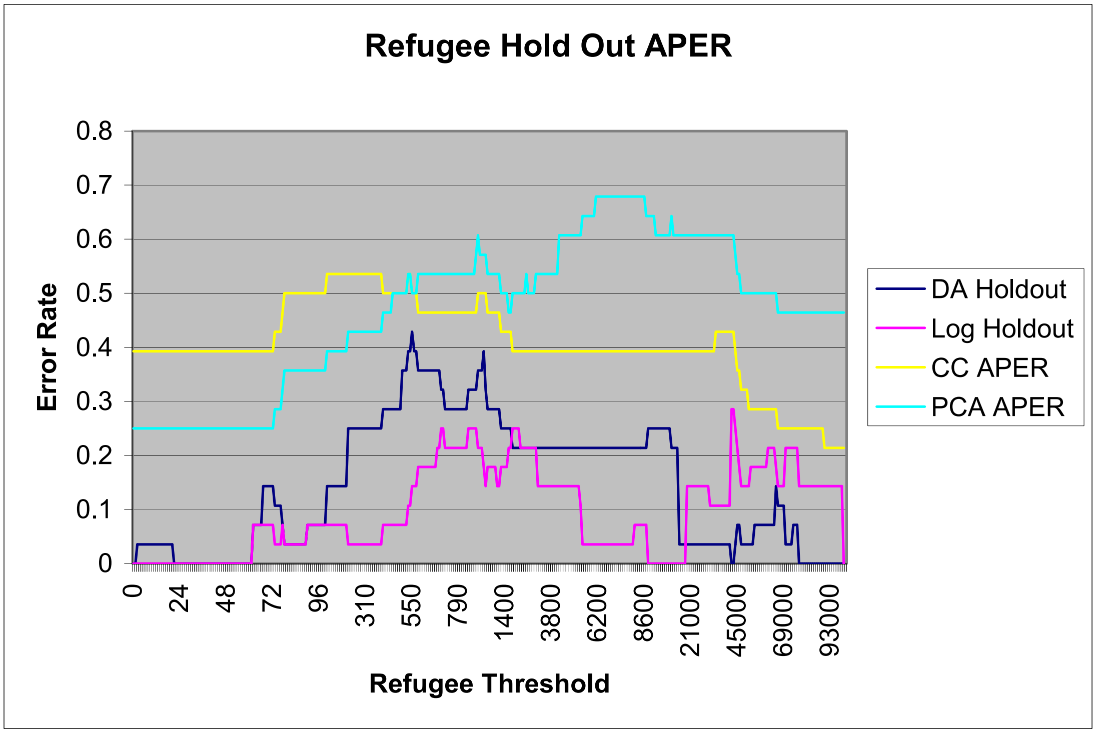


Figure Z-3. Refugee Apparent Error Rates for Several Models

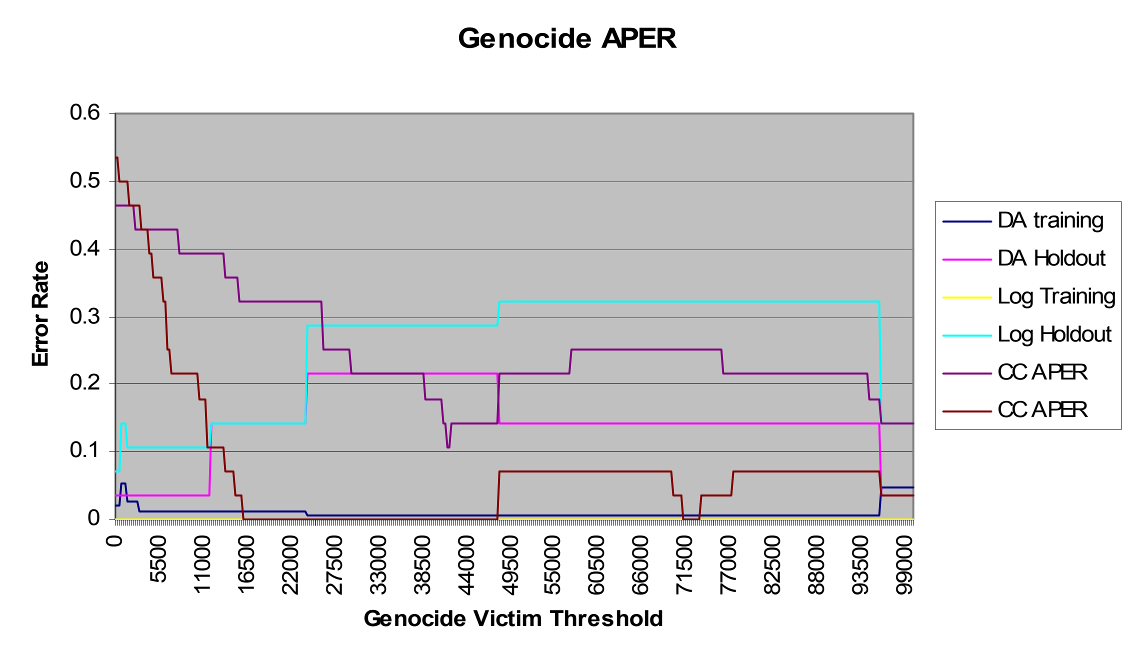


Figure Z-4. Genocide / Politicide Death Apparent Error Rate for Several Models

Z.2. Principal Component Analysis Scores

The same 54 variables were used for PCA analysis as were used for canonical correlation, and are listed in Appendix K. The raw data was handled slightly differently for PCA than canonical correlation, however. There is no built in ability to convert hold out raw data into PCA scores the way that raw hold out data can be converted to canonical correlation scores using the transforming A matrix. Thus, PCA was used on the entire set of entering data from 1975 through 2002. The justification for this comes from how hold out data would be used if a researcher in 2002 wished to build a predictive model similar to the one used in this study. They would likely use all the data available up to 2002 in order to provide a more complete model of the data's structure. After PCA loadings and scores were found for the nearly complete data set, the scores were separated into their normal training and hold out sets. The variances were examined and 13 components retained based on a scree line test, and the consideration for which eigenvalues exceeded 1. Figure Z-5 shows the eigenvalue plot. Using 13 components accounted for 88.45% of the total variance in the data. The number of components included in both the training and hold out scores were then reduced to 13 and regressed against each individual dependent variable in the training data set. The resulting models basic statistics are shown in Table Z-3.

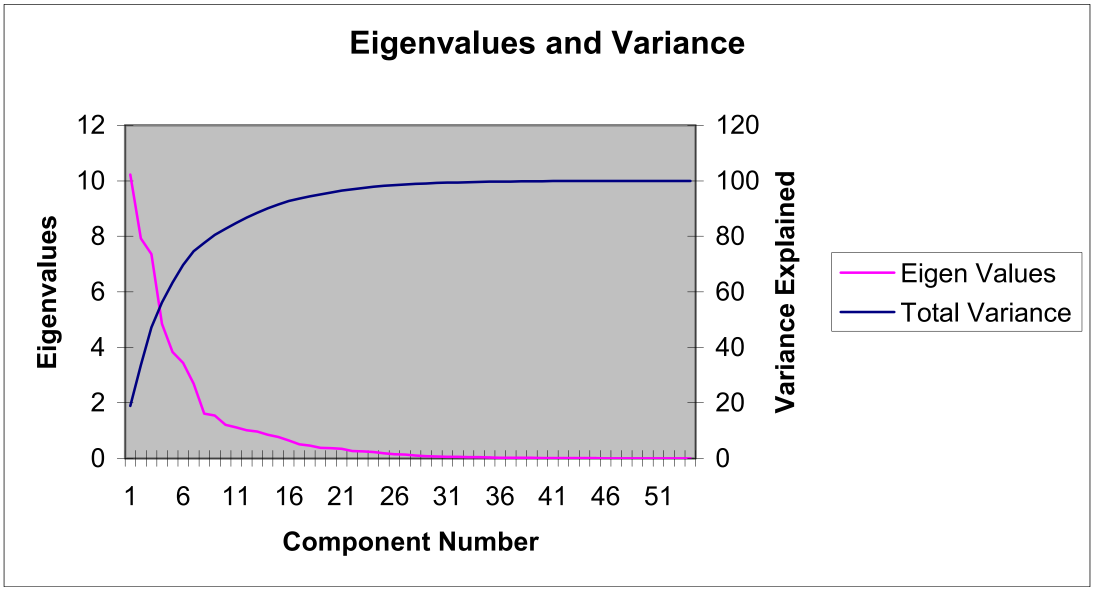


Figure Z-5. Eigenvalues and Variance Explained Using PCA

	Battle Deaths	Refugees	Genocides	Malnutrition
Rsquared	0.3293	0.67	0.471	0.8318
Rsquare Adj	0.2652	0.6385	0.4204	0.8158
Rsquare Pred	0.1381	0.6018	0.3604	0.7918
RMSE	4864.3	262990	36344	6.646

Table Z-3. PCA OLS Training Model Statistics

From each of these predictive models built based on training data a set of predictions for each of the 4 dependent variables of interest was generated using the hold out data the same way it was done using canonical correlation scores. Figure 4.9 shows the results of these predictions using OLS on PCA scores. Note that for battle deaths, refugees, and genocides the RMSE of the PCA predictions is lower, and the R^2 is higher. The exception to this is malnutrition predictions, where canonical correlation has a lower RMSE and higher R^2.

The forecasts using PCA scores suffer the same defects as using canonical correlate scores, in particular a high RMSE which limits its utility as a forecasting tool for battle deaths, refugees, and genocide. GLS models were also tried on PCA scores, yet proved no better than using the canonical correlate scores. Appendix Q shows the best results of GLS models using PCA scores.

		Battle Deaths		Refugees		Genocide		Malnutrition	
Country	Year	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Yemen	2003	0	-931.2	550	50535.2	0	11435.0	37.0	30.0
Yemen	2004	0	-1590.5	526	13193.5	0	5898.6	38.0	30.3
Yemen	2005	0	-2006.4	395	25383.9	0	6490.1	30.5	29.1
Yemen	2006	0	-2860.4	573	20332.3	0	6954.5	36.6	27.1
Somalia	2003	0	64.1	315114	334466.8	0	-16817.2	16.1	27.7
Somalia	2004	0	-626.1	306200	247116.5	0	-18429.6	19.7	24.2
Somalia	2005	0	-468.2	314066	241634.8	0	-17258.6	12.7	23.3
Somalia	2006	547	-984.0	388046	255110.6	0	-23037.3	18.6	22.9
Sudan	2003	3225	1952.4	580727	431016.5	96000	77308.2	27.0	27.9
Sudan	2004	6569	1155.3	696657	433984.2	192000	73261.5	26.0	29.4
Sudan	2005	1204	1268.9	655732	433383.5	48000	71112.5	29.2	28.7
Sudan	2006	1002	737.8	645093	455067.5	48000	69860.0	32.3	28.1
Djibouti	2003	0	-467.9	78	-509091.6	0	15968.1	26.0	37.2
Djibouti	2004	0	-167.4	73	-572391.1	0	11385.2	24.0	33.8
Djibouti	2005	0	-273.5	77	-524185.8	0	10492.0	25.2	35.0
Djibouti	2006	0	-751.4	105	-724538.8	0	1690.9	27.5	26.5
Kenya	2003	100	211.4	883	-71847.2	0	-812.7	31.0	33.2
Kenya	2004	52	442.7	935	-61552.8	0	436.3	31.0	33.2
Kenya	2005	251	-1218.7	1186	-66449.3	0	-1387.2	28.2	31.4
Kenya	2006	567	-1990.6	1753	-151067.4	0	-9488.5	27.6	32.1
Ethiopia	2003	970	10667.7	43676	350914.7	0	4247.9	46.0	55.9
Ethiopia	2004	936	9011.0	44219	120684.1	0	-1303.6	46.0	51.2
Ethiopia	2005	773	7859.4	46824	-1463.6	0	-1631.1	48.7	48.9
Ethiopia	2006	0	6989.7	65361	-65971.1	0	3174.7	48.4	48.5
Eritrea	2003	57	8837.5	116964	578394.2	0	13613.1	73.0	71.7
Eritrea	2004	0	6669.7	121937	262339.7	0	14903.6	75.0	64.9
Eritrea	2005	0	5819.5	130544	265495.5	0	4637.9	72.3	62.4
Eritrea	2006	0	5286.1	168682	189631.4	0	6453.5	71.7	61.5
Rsquared		0.4837		0.6198		0.5300		0.8074	
RMSE		4389.81		277528.37		26501.65		7.09	

Table Z-4. Predictions on Hold Our Data Using PCA and OLS

Z.3. PCA and Canonical Correlations Loadings

Despite the lack of precise predictive ability of the battle deaths, genocide, and refugee models, the use of canonical correlate scores did provide strong predictive ability of undernousrishment and thus this section will examine the PCA and canonical correlations loadings involved in each model. Appendix N shows the loadings for each of the canonical variables as it relates to predicting undernourishment. Unlike the actual data runs shown above, the data for this section was standardized prior to use in order to make the relative values of the loadings more easily interpretable. This does not alter the relationships between the variables, the models, or conclusions about them. The use of raw data was necessary to utilize hold out data. The figures and data shown in this section all pertain to the training data set.

Table Z-5 below shows the variables with loadings of magnitude greater than one on each canonical variate. The threshold of 1 was used since it suggests that the associated variable has more correlation with the outcome than random noise does.

CC Variable 1	
Loaded Variable	Loading
Undernourishment	-0.6527
Refugees	-0.6757
Ethiopia	-0.7525
Battle Deaths	-0.5803

CC Variable 2	
Loaded Variable	Loading
Undernourishment	-0.4166
Population Density	-0.4759
Agriculture as a % of GDP	0.4552
Military as a % of GDP	-0.4096
Trade Openness	-0.555
Telephones per 100	-0.4308
Forrested Land	0.4769

CC Variable 3	
Loaded Variable	Loading
Forrested Land	0.5806
Economic Discrimination	-0.6757
Pecent Paved	0.6426
Genocide Deaths	0.5729

CC Variable 4	
Loaded Variable	Loading
Anarchy	-0.2776
Partial Democracy w/ Fact	0.3616
Military as % GDP	-0.2692

Table Z-5. Significant Loadings on Canonical Variables

The 4 year lagged data on battle deaths, refugees, and undernourishment is most loaded on the variable with the greatest correlation, indicating those types of instability have persistence. Also of note is that partial democracy with factionalism is loaded on the 4th variable, which is significant only to the model of battle deaths. Table Z-6 shows the basic data from each OLS model using canonical correlation scores found using JMP, including the probability of accepting an H_0 : that the variable contributes significantly to the model.

	Battle Deaths	Refugees	Genocides	Malnutrition
R-square	0.5546	0.8979	0.7168	0.9612
R-square Adjusted	0.5423	0.8951	0.709	0.9601
RMSE	3839.06	141638	25751.6	3.0925
Variable 1 Coefficient	2117.77	193518	-8033.24	14.979
Variable 2 Coefficient	2484.77	365973	8793.6	-2.3519
Variable 3 Coefficient	-2.75	-13255	38618.7	0.6474
Variable 4 Coefficient	-2683.28	14875	547.96	0.2723
t-ratio of Variable 1	6.73	16.68	-3.81	59.13
t-ratio of Variable 2	7.9	31.54	4.17	-9.28
t-ratio of Variable 3	-0.01	-1.14	18.31	2.56
t-ratio of Variable 4	-8.53	1.28	0.26	1.07
Prob 1 >|t|	<.0001	<.0001	0.0002	<.0001
Prob 2 >|t|	<.0001	<.0001	<.0001	<.0001
Prob 3 >|t|	0.993	0.2552	<.0001	0.0116
Prob 4 >|t|	<.0001	0.2019	0.7954	0.2842

Table Z-6. Canonical Correlation Training Model Statistics

From the figure above, it can be seen that canonical variable 1 contributes to each instability indicator model, which is not unexpected due to canonical variable 1 capturing the most correlation by design, and encapsulating the lagged dependent variables. Scores from canonical variable 4 contribute significantly only to the model of battle deaths.

PCA universally yielded worse predictions than canonical correlation when developing continuous models, however this section will also provide a brief discussion of PCA loadings with regard to malnutrition as well. For all of the data shown, MI 1 was used to generate the analysis. Appendix O shows the loadings matrix for the PCA scores. Appendix P shows the results of the OLS model developed using the PCA score training data.

The strongest of the four models is the one for malnutrition. PCA variables 4 and 6 have the greatest influence on the malnutrition model, where variable 4 has a positive correlation with malnutrition and variable 6 a negative one. Variable 6 has year, anarchy, literacy, and Somalia data positively loaded on it. However, the canonical correlation model surpasses using PCA in terms of both R^2 and RMSE.

Z.4. Individual Country Models

The previous results have shown that there are influences not captured by the individual variables in the raw data that are instead captured by the catch all binary designators of country, or by variables within the data that act as proxies for identifying an individual country (such as Yemen’s religious fractionalization score of 0). This in turn indicated that different variables may have different effects on different countries in the Horn of Africa for continuous forecasting purposes. Based on the results of whole region models, canonical correlation scores were used to generate individual country forecasts for battle deaths, refugees, and malnutrition. PCA scores were used to generate predictions for genocide. In general, where canonical correlation and PCA performed similarly canonical correlation was preferred since it used only 4 variables vice 13 with the greatly reduced dataset, which had the effect of reducing variance in the model and its predictions. In addition, genocide models were only generated for those countries in the region which have experienced genocide during the training set time frame (Somalia, Sudan, and Ethiopia). The results of the pilot study can be found in Appendix R.

These models suffered from a loss of degrees of freedom, particularly in the case of Eritrea. As a whole, they did not improve on the RMSE of predictions for the countries as a regional whole. The training data models fit very well, to the point that many of them had nearly zero residual and variance. However, they did not provide better predictions when hold out data was applied to the by country models. It is possible that with more observations accurate models built on individual counties could be created which would provide better forecasts, however given the data set available for this study further models will continue to predict instability indicators for each country using all regional data available to reduce variance. This makes sense, since the primary failing of the previously discussed regional models for battle deaths, refugees, and genocide were they contained too much variance. By reducing the number of degrees of freedom, variance was inflated even further.

Z.5. Conclusions on Continuous Models

Many of the continuous models examined in Appendix Z demonstrated promising $R^2_{\Pr ed}$ and RMSE when used on the training data. However, when applied to the hold out data, only models of malnutrition had a small enough variance to offer significant potential utility for forecasters examining countries in the Horn of Africa. However, the use of canonical correlation and advanced regression techniques appears unnecessary for forecasting malnutrition, since OLS regression produced results that were not statistically different from those achieved using more complicated methods. The variables most heavily weighted on the 4 year canonical correlation forecasting model of malnutrition, in decreasing order of importance based on *t*-testing, were ethnic fractionalization, anarchy, linguistic fractionalization, urban population, and telephone subscribers. Of these variables, only anarchy and urban population increase malnutrition. Using a mixed stepwise regression, the variables most significant to the model, with $p < .05$, were year, literacy, population, full autocracy, years since last conflict, water per capita, population density, arable land per capita, and the water per capita / arable land per capita interaction.

Based on these results, the focus of this study shifted towards discrete prediction models for battle deaths, refugees, and genocides. The initial assessment that an exact, continuous estimate is preferable to a discrete one stands, and it was judged that the $R^2_{\Pr ed}$ and RMSE of the malnutrition hold data indicates sufficient utility to render it of more utility than a simpler over / under prediction gained via DA or logistic regression, no matter the error rate of the other discrete classification methods. Further analysis in this study therefore focused on discrete models forecasting battle deaths, refugees, and genocide.

Bibliography

Adebayo, Adejei, Afari-Gyan, K., Chin, Dato, *et al*, *Kenya General Election 27 December 2002,* Commonwealth Observer Group, London, U.K. 2006

Alesina, Alberto, Devleeschauwer, Amaund, Easterly, William, and Kurlat, Sergio,. "Fractionalization" - *Journal of Economic Growth,* vol. 8, no. 2, June 2003.

Allard, Kenneth. *Somalia Operations: Lessons Learned,* National Defense University Press. Washington D.C. 1995

Allen, D.M. "Mean Square Error of prediction as a criterion for selecting variables". *Technometrics*, 1971

Artelli, Michael J. "Public Resolve: The Casualty of the Long War". Presented to the Military Operations Research Symposium, May 2007

Artz, Donna E., *Refugees into Citizens: Palestinians and the End of the Arab-Israeli Conflict*, Council on Foreign Relations Press, Washington D.C. 1997

Allison, P.D. *Missing Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage, 2001.

Banks, Jerry, *et al,. Discrete- Even System Simulation,* Prentice Hall, Upper Saddle River NJ, 2005

Barnett, Thomas P.M. "The Pentagon's New Map," *Esquire*, March 2003.

BBC News, *Flashback 1948: Portrait of a Famine*, April 6th 2000 http://news.bbc.co.uk/1/hi/world/africa/703958.stm accessed 9 November 07

BBC News, *Profile: Somalia's Islamic Courts*, June 6th 2006, http://news.bbc.co.uk/1/hi/world/africa/5051588.stm, accessed January 12, 2008

Box, G.E.P. and Cox, D.R. "An Analysis of Transformations," *J.R. Stat. Soc. Ser. B*, 1964

Burden, Richard L. and Faires, Douglas J. *Numerical Analysis.* Pacific Grove: Brooks/ Cole, 2001.

Cattell, R. B., "The Screen Test for the Number of Factors", *Multivariate Behavioral Research*, vol. 1, 140-161, 1966

Cleveland, W.S. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829- 836. 1979

Cleveland, W.S. and Devlin, S.J. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, Vol. 83, pp. 596-610. 1988

Center for Army Analysis (CAA), "Analyzing Complex Threats for Operations and Readiness", Fort Belvoir VA, September 2001.

Cho, Namsuk, "Critical Infrastructure Rebuild Prioritization Using Simulation Optimization", MS Thesis, AFIT/ENS/07M-04. School of Operations Research, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2007.

Collier, Paul and Anke Hoeffler "On economic causes of civil war," *Oxford Economic Papers 50*, 1998.

Collier, Paul and Hoeffler, Anke. "On the Incidence of Civil War in Africa", *Journal of Conflict Resolution,* 2002.

Collier, Paul and Hoeffler, Anke, "Greed and Grievance in Civil War", *Oxford Economic Papers Vol 56 Number 4*, August 2004.

Congressional Budget Office (CBO), "Preliminary Analysis of the President's FY 2008 Budget request", March 2006. http://www.cbo.gov/ftpdocs/78xx/doc7836/03-02-Prelim_Analysis.pdf

CNN, "*Truck bomb kills chief U.N. envoy to Iraq*", August 20th 2003, http://www.cnn.com/2003/WORLD/meast/08/19/sprj.irq.main/index.html Accessed February 12th 2008

de Waal, Alex and Flint, Julie. *Darfur: A Short History of A Long War.* Zed Books, New York. 2006.

Department of Defense Directive Number 3000.05, Military Support for Stability, Security, Transition, and Reconstruction (SSTR) Operations. 28 November 2005.

Durch, William J. "Are There Peace/Stability Operations in Your Future?". Briefing for the Chairman of the Joint Chiefs of Staff Seminar, 12 June 2002.

Durbin,J. and Watson, G.S. "Testing for Serial Correlation in Least Squares Regression II," *Biometrica* 1951

Enders, Walter. *Applied Econometric Time Series: Second Edition,* New York, John Wiley and Sons, 2004

Esty, Daniel C. *et al*. Phase II State Failure Task Force Report. Political Instability/State Failure Task Force Website, http://globalpolicy.gmu.edu/pitf/SFTF%20Phase%20I%20Report.pdf, accessed 21 October 2007

Evans, Graham and Jeffrey Newnham. *The Dictionary of World Politics. A Reference Guide to Concepts, Ideas and Institutions*. London, Harvester Wheatsheaf, 1992.

Forest, James J. F. *The Making of a Terrorist: Volume Three, Root Causes*. Praeger Security International, Connecticut, 2006.

GlobalSecurity.org, "Ethiopia / Eritrea War," 2005, http://www.globalsecurity.org/military/world/war/eritrea.htm accessed 29 JAN 08

Goldstone, Jack A., Bates, Robert H., Gurr, Ted R., Lustik, Michael, Marshall, Monty G., Ulfelder, Jay, Woodward, Mark, Political Instability Task Force. "A Global Forecasting Model of Political Instability." Paper prepared for presentation at Annual Meeting of the American Political Science Association, September 2005.

Gurr, Ted Robert and Barbara Harff. *Early Warning of Communal Conflicts and Genocide: Linking Empirical Research to International Responses*. Tokyo, The United Nations University, 1996.

Gurr, Ted Robert and Harff, Barbara., Systematic Early Warning of Humanitarian Emergencies. *Journal of Peace Research,* 551-571, 1998

Gurr, Ted Robert, Woodward, Mark, and Marshall, Monty G., "Forecasting Instability: Are Ethnic Wars and Muslims Different?", Presented to Annual Meeting of the American Political Science Association, September 2005

Harmony Project. *al-Qaida's (Mis)Adventures in the Horn of Africa*. Combating Terrorism Center at West Point, 2007

Horton, Nicholas J. and Kleinman, Ken P., "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models". *The American Statistician, Vol 61 No. 1* February 2007

Honaker, James, and King, Gary, "What to do about Missing Values in Time Series Cross-Section Data", Working Papers, September 2007, accessed 24 NOV 07 http://gking.harvard.edu/files/pr.pdf

Jaquin-Berdal, Dominique. *Nationalism and Ethnicity in the Horn of Africa: A Critique of the Ethnic Interpretation.* The Edwin Mellen Press, 2002

JMP. Version 6.0.2. IBM Computer Software. S.A.S. Institute, Cary NC, 2006

Johnson, Richard A. and Wichern, Dean W. Applied *Multivariate Statistical Analysis.* Upper Saddle River: Prentice Hall, 2002.

Kaiser, H.F., "The Varimax Rotation for Analytic Rotation in Factor Analysis", *Psychometrica*, vol 23, 187-200, 1958

King, Gary, Honaker, James, Joseph, Anne, and Scheve, Kenneth. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation", *American Political Science Review, Vol. 95 No. 1* March 2001

Lattin, James M., J. Douglas Carroll, Paul E. Green. *Analyzing Multivariate Data*. Duxbury, 2003.

Law, Averil M., *Simulation and Modeling Analysis*, New York , McGraw-Hill Education, 2007

Little, Roderick J.A. and Rubin, Donald B., *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987

Lophaven, S.N., Nielsen, H.B., Sondergaard, J., *DACE: A MATLAB Kriging Toolbox,* Technical University of Denmark, 2002

Makridakis, Spyros, Wheelwright, Steven C. and McGee, Victor T., *Forecasting: Methods and Applications*. York: John Wiley & Sons, 1983

Mardia, K.V., Kent, Jt., and Bibby, J.M. *Multivariate Analysis.* London: Academic Press Inc. 1979.

Marshall, Monty G., Director, Polity IV and Armed Conflict and Intervention Projects Research Director, Center for Global Policy Research Professor, School of Public Policy George Mason University, Arlington, VA, Personal Electronic correspondence, 18 September 2007

Minorities At Risk Project. "Dataset Users Manual 030703", http://www.cidcm.umd.edu/mar/margene/mar-codebook_040903.pdf accessed 21 October 2007

Montgomery, Douglas C., Elizabeth A. Pec and G. Geoffrey Vining. *Introduction to Linear Regression Analysis.* New York: John Wiley & Sons, 2006.

Montgomery, Douglas C. *Design and Analysis of Experiments*. New York: John Wiley & Sons, 2005.

Nafziger, E. Wayne and Juha Auvinen. *War, Hunger, and Displacement: An Econometric Investigation into the Sources of Humanitarian Emergencies*. Helsinki: UNU World Institute for Development Economics Research, 1997.

Nau, Robert F. Introduction to Forecasting 411 Notes. 2005. http://www.duke.edu/~rnau/411arim.htm, accessed 19 November 07

Niskanen, William, A. "Crime, Police, and Root Causes". Cato Policy Analysis Paper No. 218, November 1994. http://www.cato.org/pubs/pas/pa-218.html, accessed 21 October 2007.

Nysether, Nathan E., "Classifying Failing States", MS Thesis, AFIT/ENS/07M-19. School of Operations Research, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2007.

O'Brien, Sean P. Analyzing Complex Threats for Operations and Readiness (ACTOR): A Methodology for Forecasting Country Instability. Presented to the Military Operations Research Society Symposium, July 2002.

Pate, Amy. Minorities at Risk Director, University of Maryland, College park, MD, Personal electronic correspondence, 6 November 2007

Pfetsch, Frank R. and Rohloff, Cristoph. "KOSIMO: A Databank on Political Conflict", Journal of Peace Research Vol. 37 No. 3, 2000.

Power, Daniel J. "Ask Dan! - What is the "true story" about data mining, beer and diapers?" DSS News, Vol 3. No. 23, November 10th 2002, http://www.dssresources.com/newsletters/66.php

Robbins, Matthew JD. "Investigating the Complexities of Nationbuilding: A Sub-National Regional Perspective", MS Thesis, AFIT/ENS/05M-16. School of Operations Research, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2007.

Rotberg, Robert I., "When States Fail: Causes and Consequences", MIT Security Studies Seminar Fall 2003. http://web.mit.edu/ssp/seminars/wed_archives_03fall/rotberg.htm

Rubin, Donald B. Multiple Imputation for Non-Response in Surveys. New York: New York: John Wiley & Sons, 1987

Schafer, J.L. *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman and Hall, 1999

Schmid, Alex P. “Thesaurus and Glossary of Early Warning and Conflict Prevention Terms.” Abridged version edited by Sanam B. Anderlini for FEWER. Rodderdam: Synthesis Foundation, 1998.

Schatzman, Michelle. *Numerical Analysis*. Oxford: Clarendon Press, 2002

Stevens, James G. “An Investigation of Multivariate Adaptive Regression Splines for Modeling and Analysis of Univariate and Semi-Multivariate Time Series Systems.” PhD. Dissertation, School of Mathematics, Naval Post Graduate School, Monterey CA, 1991

Takeyh, Ray and Nikolas Gvosdev. Do Terrorist Networks Need a Home? The Washington Quarterly, 97-108, Summer 2002.

Tareke, Gebru "The Ethiopia-Somalia War of 1977 Revisited," in Board of Trustees, Boston University, *The International Journal of African Historical Studies*. Boston University African Studies Center, 2000

Ulfelder, Jay, Research Director Political Instability Task Force, Science Applications International Corporation (SAIC), personal electronic correspondence, 10 October 2007

United Nations Department of Peacekeeping operations (UNDPKO). “Somalia – UNOSOM I: Background”, http://www.un.org/Depts/dpko/dpko/co_mission/unosom1backgr2.html accessed 29 JAN 08

United Nations Food and Agriculture Organization (UNFAO). “Study on the Impact of Armed Conflicts on the Nutritional Situation of Children” 1995 http://www.fao.org/docrep/005/w2357e/W2357E11.htm

United Nations High Commissioner of Refugees (UNHCR). “Convention and Protocol Relating to the Status of Refugees” 1951, 1967 http://www.unhcr.org/cgi-in/texis/vtx/protect/opendoc.pdf?tbl=PROTECTION&id=3b66c2aa10

USAid, “Fiscal Year 2008 Budget Request”, September 2007 http://www.usaid.gov/policy/budget/cbj2008/fy2008cbj_highlights.pdf

Wackerly, Dennis D., William Mendenhall III and Richard L. Scheaffer. *Mathematical Statistics with Applications 6th Edition*. Duxbury, 2002.

Weiss, Thomas and Cindy Collins., *Humanitarian Challenges and Intervention. World Politics and the Dilemmas of Help*, Westview Press, Boulder CO, 1996.

Wiener, Norbert. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* New York: John Wiley & Sons, 1949.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 14-03-2008	**2. REPORT TYPE** **Master's Thesis**	**3. DATES COVERED** *(From – To)* Aug 2006 – Mar 2008

4. TITLE AND SUBTITLE FORECASTING INSTABILITY INDICATORS IN THE HORN OF AFRICA REGION	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Tannehill, Bryan R.	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Street, Building 642 WPAFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT/GOR/ENS/08-21
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Edwin J.Offutt, Lt Col, USAF USCENTCOM J8-ARB 7115 South Boundary Blvd MacDill AFB, FL 33621	**10. SPONSOR/MONITOR'S ACRONYM(S)** USCENTCOM **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The forecasting of state failure and the associated indicators has been a topic of great interest to a number of different agencies. USAid, CENTCOM, the World Bank, the Center for Army Analyses, and others have all examined the subject based on their own specific objectives. Whether the goal is denying terrorists space in which to operate, deciding how to pre-position materials in anticipation of unrest, stabilizing foreign markets and trade, or preventing or mitigating humanitarian disasters, man made or otherwise, this topic has been of interest for over a decade.

The Horn of Africa has been one of the least stable regions in the world over the past three decades, and a continual source of humanitarian crises as well as terrorist activity. Some of the initial modeling of instability was done in response to crises in the Horn of Africa, but research is ongoing. Current models forecasting instability suffer from lack of lead time, subjective predictions, and lack of specificity. The models demonstrated in this study provide 4 year forecasts of battle deaths per capita, refugees per capita, genocide, and undernourishment for Djibouti, Ethiopia, Eritrea, Kenya, Somalia, Sudan, and Yemen. This thesis used principal component analysis, canonical correlation, ordinary least squares regression, logistic regression, and discriminant analysis to develop models of each instability indicator using 54 variables covering 32 years of observations. The key variables within each model are identified, and the accuracy of each model is compared with current models.

15. SUBJECT TERMS

Failing States, Horn of Africa, Missing Data, Principal Component Analysis, Canonical Correlation, Discriminant Analysis, Logistic Regression

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Dr Richard F. Deckro, Professor of Operations Research, (ENS)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4325; e-mail: richard.deckro@afit.edu
U	U	U	UU	270	